Award Number: DAMD17-02-1-0373

TITLE: Modular Machine Learning Methods for Computer-Aided Diagnosis of Breast Cancer

PRINCIPAL INVESTIGATOR: Jonathan L. Jesneck
Joseph Lo, Ph.D.

CONTRACTING ORGANIZATION: Duke University
Durham, North Carolina 27710

REPORT DATE: May 2003

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 2003	Annual Summary (1 May 02 - 30 Apr 03)

4. TITLE AND SUBTITLE
Modular Machine Learning Methods for Computer-Aided Diagnosis of Breast Cancer

5. FUNDING NUMBERS
DAMD17-02-1-0373

6. AUTHOR(S)
Jonathan L. Jesneck
Joseph Lo, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Duke University
Durham, North Carolina 27710

E-Mail: jonathan.jesneck@duke.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

The purpose of this study was four fold: 1) Identify subsets of the training data of breast cancer features using both a priori information and unsupervised learning methods. 2) Build local models for breast cancer prediction for each subset of the training data using supervised learning methods. Evaluate the performance of the local models on the training data relative to a single, global, supervised learning model and to current clinical practice. 3) Combine the local models to form a global, modular model and evaluate the performance of the modular model on the evaluation data set relative to a single, global, supervised learning model and to current clinical practice. 4) Develop an ensemble classifier combining three sources of data (image processing, radiologist-extracted mammographic findings, and patient history) for the task of computer-aided diagnosis of breast microcalcification clusters. We developed the world's largest database of over 4400 cases containing radiologist-extracted mammographic findings, patient age, and biopsy outcome, and we used this data to develop modular, global CAD models using different machine learning algorithms applied to the entire database.

14. SUBJECT TERMS
computer-aided diagnosis, breast cancer, BI-RADS, image processing, ensemble classifier

15. NUMBER OF PAGES
45

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Cover.....1

SF 298.....2

Table of Contents.....3

Introduction.....4

Body.....4

Key Research Accomplishments.....7

Conclusions.....7

Reportable Outcomes.....7

References.....8

Appendices.....8

INTRODUCTION

This progress report covers projects by two different students both mentored by Joseph Lo. The predoctoral fellowship was originally awarded to Mia Markey, who graduated in 2002 from Duke University with her Ph.D. *Modular Machine Learning Methods for Computer-Aided Diagnosis of Breast Cancer*. The Army authorized the transfer of the remaining fellowship to Jonathan Jesneck who is currently iin his first year of graduate studies. His research to date has continued work in modular models combining data from different sources to create predictive models for breast cancer.

BODY – Part 1: Dissertation Research by Mia Markey

This study investigated modular and ensemble systems of machine learning methods for computer-aided diagnosis (CAD) of breast cancer to reduce the number of benign biopsies. While mammography is valuable for early detection of breast cancer, it has a high false-positive rate. A CAD system for referring benign lesions to short-term follow-up instead of biopsy could spare women discomfort, anxiety, and expense and potentially improve the cost-effectiveness of mammographic screening programs.

We began with an analysis of the error surfaces mapping the internal parameters of a simple Perceptron model to several measures of performance, and surmised that the default behavior of minimizing the mean squared error may also maximize the ROC area under the curve (AUC), but may not maximize the partial area (partial AUC) [1] under the high sensitivity region of the ROC curve. The partial AUC is more clinically relevant than the AUC since a breast cancer CAD system must maintain high sensitivity (i.e., delaying the diagnosis of breast cancer is generally worse than a benign biopsy). This work was published by the fellow and mentor in the peer-reviewed journal *Computers in Biology and Medicine* (App. 1).

Task 1. Identify subsets of the training data using both a priori information and unsupervised learning methods.

This database is to our knowledge the largest of its kind, consisting of 4435 cases from 3 sources: Duke, University of Pennsylvania, and the Army-sponsored DDSM. The database was randomly split into two halves for training vs. evaluation. Each biopsy-proven breast lesion was described by six mammographic findings and patient age.

We clustered the data according to institution, lesion type, and patient age. We further studied clustering using three unsupervised learning techniques: agglomerative hierarchical clustering followed by K-Means [2], Self-Organizing Map [3], and AutoClass [4]. All three identified a cluster of mostly benign cases, which may be candidates for obviating biopsy. Using profiling techniques, we described a typical case in these clusters as a younger woman with a well-circumscribed or obscured, oval-shaped mass. In other words, each clustering technique identified a cluster that contained the majority of the benign cases that would have been correctly referred to follow-up.

Task 2. Build local models for breast cancer prediction for each subset of the training data using supervised learning methods. Evaluate the performance of the local models on the training data relative to a single, global, supervised learning model and to current clinical practice.

To establish the best single global model for predicting biopsy outcome from mammographic findings and patient age, we exhaustively investigated five supervised machine learning methods: linear discriminant analysis (LDA) [2], support vector machines (SVM) [5], back-propagation artificial neural networks (BP-ANN) [6], case-based reasoning (CBR) [7], and

classification and regression trees (CART) [8]. The non-linear models (BP-ANN, CBR, CART) were superior to the linear models (LDA, SVM) for this task. In our extensive previous work with smaller data sets, we had not been able to demonstrate this expected superiority of non-linear models over linear models for this task. The global BP-ANN and global CBR models peformed the best with the highest partial AUC values. For the BP-ANN, the AUC = 0.820 ± 0.009 and the partial AUC = 0.347 ± 0.022. For the CBR, the AUC = 0.788 ± 0.009 and the partial AUC = 0.324 ± 0.019.

We examined the performance of the global models over the local subsets identified using a priori knowledge and unsupervised learning. One of the most striking results of this dissertation was that CAD systems trained on a mixture of lesion types performed much better on masses than on calcifications. This work was published by the fellow and mentor in the peer-reviewed journal *Radiology* (App. 2). The study of the institutional effects suggests that models built on cases mixed between institutions may overcome some of the weaknesses of models built on cases from a single institution. The same threshold on the BP-ANN gave approximately 98% sensitivity and 23% specificity on all three of the institution subsets. There was no benefit in training a BP-ANN specifically for cases at each institution compared to the global BP-ANN trained on cases from all institutions. We also found that a BP-ANN trained on cases from one institution did not always perform the same way (AUC, partial AUC, sensitivity and specificity for a fixed threshold) when tested on cases from another institution. Thus, further cross-institutional studies of breast cancer CAD systems are still needed.

Task 3. Combine the local models to form a global, modular model. Evaluate the performance of the modular model on the evaluation data set relative to a single, global, supervised learning model and to current clinical practice.

We then developed modular CAD systems by building local models specifically for each of the clusters identified using a priori knowledge and unsupervised learning. While some local models were superior to some global models, we were unable to build a modular CAD system that was better than the global BP-ANN model, which was considered to be a "gold standard" since we have used BP-ANN models extensively in our laboratory and the overall performance of the global BP-ANN was generally better than that of the other global models. We consider it unlikely that additional work with similar modular systems would prove fruitful. However, the cluster analysis and local models also led to an unexpected, interesting result. We developed a simple diagnostic rule from the local CART model for masses and the profiles of the very likely benign mass clusters identified by the unsupervised learning methods: if the Mass Margin was well-circumscribed or obscured and the age was less than 59 years and there were no calcifications, associated findings, or special findings, then don't biopsy, otherwise do biopsy. On the 2258 training cases, this clinically intuitive rule gave 961 / 982 = 98% sensitivity and 336 / 1276 = 26% specificity. In other words, this simple rule performed comparably to the complicated global BP-ANN machine learning model with a threshold of 0.1842 (965 / 982 = 98% sensitivity, 303 / 1276 = 24% specificity). Compared to more complicated models, such a simple rule would be trivial to implement, make it more understandable and thus perhaps acceptable to clinicians, and allow for comparisons to existing clinical criteria. This clustering work was published by the fellow and mentor in the peer-reviewed journal *Artificial Intelligence in Medicine* (App. 3). Other work was also performed as part of this dissertation but cannot be described here due to length restrictions of this report. The dissertation is 178 pages and is available upon request.

BODY – Part 2: First Year Research by Jonathan Jesneck

The purpose of this research was to develop a computer aid to help radiologists identify whether suspicious calcification clusters are benign vs. malignant, such that they may potentially recommend fewer unnecessary biopsies for actually benign lesions. We developed an ensemble classifier for the task of computer-aided diagnosis of breast microcalcification clusters, which are very challenging to characterize for radiologists and computer models alike. We investigated combining these three sources of data (image processing, radiologist-extracted mammographic findings, and patient history) into one ensemble system. An ensemble system uses multiple classifiers to solve a classification problem by training multiple models for the same cases and then combining models' predictions [9]. Simple ensembles of classifiers using voting or averaging to combine their predictions have shown promise in this field [10, 11]. The hypothesis was that an ensemble classifier comprised of information from all three sources of data can significantly outperform models based upon local subsets of data.

The data consisted of mammographic features extracted by automated image processing algorithms as well as manually interpreted by radiologists according to a standardized lexicon. We used 292 cases from the Army's DDSM mammography database. From each case, we extracted 22 image processing features pertaining to lesion morphology, 5 radiologist features also pertaining to morphology, and the patient age.

Linear discriminant analysis (LDA) models were designed using each of the three data types. Each local model performed poorly; the best was one based upon image processing features which yielded AUC of 0.59 ± 0.03 and partial AUC of 0.08 ± 0.03. We then developed ensemble models using different combinations of those data types, and these models all improved performance compared to the local models. The final ensemble model was based upon 5 features selected by stepwise LDA from all 28 available features. This ensemble performed with AUC of 0.69 ± 0.03 and partial AUC of 0.21 ± 0.04, which was statistically significantly better than the model based on the image processing features alone ($p<0.001$ and $p=0.01$ for full and partial AUC respectively). This demonstrated the value of the radiologist-extracted features as a source of information for this task. It also suggested there is potential for improved performance using this ensemble classifier approach to combine different sources of currently available data. This work was presented at SPIE Medical Imaging 2003 conference and is currently in press for the proceedings (App. 4). The data are described in more detail in Table 2 and Figure 2 of that appendix.

In on-going work, we are calculating additional shape features of the clusters of calcifications to improve classifier performance. The following cluster features were added to the CAD code: mean calcification area and bounding boxes. New calcification features were first moments, second moments, moment ratio, and elliptical features. For each cluster, the average, standard deviation, minimum, maximum, and coefficient of variation were calculated over all the calcifications in that cluster, and these summary features were used as additional cluster features. In all 69 new cluster features were calculated.

The current version of the CAD algorithm requires up to one hour to process each patient case of 4 mammograms. To make the code faster, it was profiled and the most computationally intensive sections of the algorithm were optimized. Additionally, the two most time consuming functions were parallelized and evaluated on different parallel architectures, ranging from a dual processor Linux machine to a Beowulf cluster. The performance speedup was encouraging, and once the CAD algorithm is in its final state, the entire project may be parallelized.

KEY RESEARCH ACCOMPLISHMENTS

- For a simple perceptron computer-aided diagnosis (CAD) model that predicted whether a breast lesion is benign or malignant, compared the optimality of several commonly used measures of performance (mean squared error vs. full or partial ROC area).
- Developed the world's largest database of over 4400 cases containing radiologist-extracted mammographic findings, patient age, and biopsy outcome.
- Devleoped global CAD models using different machine learning algorithms applied to the entire database, and identified the BP-ANN model as the best performer.
- Identified local clusters of the data according to a priori constraints such as institution, lesion type, and age, as well as according to unsupervised learning techniques, and discovered substantial variations in performance across the different clusters, such as between mass and calcification lesion types.
- Analyzed a specific cluster of mostly benign cases common to all three unsupervised learning techniques investigated, and based upon that cluster profile, developed a simple diagnostic criterion yielding 98% sensitivity and 26% specificity for breast cancer diagnosis.
- Developed modular CAD systems by optimizing local models specifically for each of the clusters, then combining those local models to form a global, modular model. Although some local models were superior to some global models, we were unable to build a modular CAD system that was better than the global BP-ANN model.
- Continued the modular/ensemble approach by focusing on the CAD of calcification clusters, identified as the most challenging subset in the previous work. Specifically, developed ensemble models by combining three different sources of input data: radiologist findings and patient age as before, plus image processing features describing lesion morphology. Demonstrated significant improvements in ensemble model performance over the use of image processing features alone.

CONCLUSIONS

We investigated machine learning techniques for the computer-aided diagnosis of breast cancer. In particular, the goal was to increase the specificity of mammography-induced breast biopsy. This is a timely and significant problem in biomedical engineering. The largest data set of its type was assembled from three independent institutions, including the Army's DDSM. A wide variety of modeling techniques were evaluated, individually and in tandem with each other. The data were likewise analyzed as a global whole and in terms of subsets. The overall intent was to engineer modular and ensemble systems using this large data set and the rich variety of tools available. Somewhat to our surprise, these systems tended to match but not exceed the performance of a classic feed-forward, back-propagation artificial neural network. As a result of this endeavor, however, we clearly identified both the potential promises and problems inherent in the use of a large, heterogeneous data set, e.g., issues such as generalization across institutions and important difference between subtypes of cases such as masses vs. calcifications. As a result of that last discovery, we are focusing efforts now on developing modular/ensemble models for calcifications.

REPORTABLE OUTCOMES

The following publications are attached as appendices 1-4 with the same numbers. The fellow and mentors' names are boldfaced for emphasis.

1 **Markey MK, Lo JY**, Vargas-Voracek R, Tourassi GD, and Floyd CE, Jr, "Perceptron error surface analysis: A case study in breast cancer diagnosis," *Computers in Biology & Medicine* 32, 99-109 (2002).

2 **Markey MK, Lo JY**, and Floyd CE, Jr, "Differences between computer-aided diagnosis of breast masses and that of calcifications," *Radiology* 223, 489-493 (2002).

3 **Markey MK, Lo JY**, Tourassi GD, and Floyd CE, Jr, "Self-organizing map for cluster analysis of a breast cancer database," *Artificial Intelligence in Medicine* 27, 113-127 (2003).

4 **Lo JY**, Gavrielides MA, Markey MK, and **Jesneck JL**, "Computer-aided classification of breast microcalcification clusters: Merging of features from image processing and radiologists," in SPIE Medical Imaging 2003: Image Processing, (2003).

Mia Markey, the original recipient of this predoctoral fellowship, received her Ph.D. in biomedical engineering from Duke University in 2002. Her dissertation (178 pages) is entitled *Modular Machine Learning Methods for Computer-Aided Diagnosis of Breast Cancer*. She applied for over 30 faculty positions and received multiple offers. She is now an assistant professor of biomedical engineering at University of Texas at Austin, where she is continuing breast cancer research based upon her Ph.D. work. The mentor Joseph Lo has applied for NIH R01 funding based on both aspects of the research described herein; those two proposals are still pending. The second fellow Jonathan Jesneck has applied for pre-doctoral fellowships from the NSF and Whitaker Foundation but unfortunately those have not been accepted.

REFRENCES

1. Jiang Y, Metz CE, and Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology 1996; 201: 745-750.
2. Sharma S, *Applied Multivariate Techniques*. 1996, John Wiley & Sons, Inc.
3. Kohonen T, *Self-Organizing Maps*. Springer Series in Information Sciences, ed. T.S. Huang, T. Kohonen, andM.R. Schroeder. 1995, Springer-Verlag.
4. Cheeseman P and Stutz J, *Bayesian Classification (AutoClass): Theory and Results,* in *Advances in Knowledge Discovery and Data Mining,* U.M. Fayad, *et al.*, Editor. 1996, AAAI Press/MIT Press:
5. Cristianini N and Shawe-Taylor J, *An introduction to support vector machines: and other kernel-based learning methods*. 2000, Cambridge, United Kingdom: Cambridge University Press.
6. Bishop CM, *Neural Networks for Pattern Recognition*. 1995, Oxford University Press.
7. Floyd CE, Jr, Lo JY, and Tourassi GD. Cased-based reasoning computer algorithm that uses mammographic findings for breast biopsy decisions. AJR. American Journal of Roentgenology 2000; 175: 1347-1352.
8. Chambers JM and Hastie TJ, ed. *Statistical Models in S*. 1992, Wadsworth & Books/Cole Advanced Books & Software: Pacific Grove, California. 608.
9. Sharkey AJC, ed. *Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems*. Perspectives in Neural Computing, ed. J.G. Taylor. 1999, Springer-Verlag.
10. Li L, Zheng Y, Zheng L, and Clark RA. False-positive reduction in CAD mass detection using a comptetive classification strategy. Medical Physics 2001; 28: 250-258.
11. Zheng B, Chang YH, and Gur D. Mass detection in digitized mammograms using two independent computer-assisted diagnosis schemes. AJR. American Journal of Roentgenology 1996; 167: 1421-4.

APPENDICES

Four publications are attached, see "Reportable Outcomes" above for the list.

PERGAMON

Computers in Biology and Medicine 32 (2002) 99–109

Computers in Biology and Medicine

www.elsevier.com/locate/compbiomed

Perceptron error surface analysis: a case study in breast cancer diagnosis ☆

Mia K. Markey[a,b,*], Joseph Y. Lo[a,b], Rene Vargas-Voracek[b], Georgia D. Tourassi[b], Carey E. Floyd Jr.[a,b]

[a] *Department of Biomedical Engineering, Duke University, Durham, NC 27708, USA*
[b] *Digital Imaging Research Division, Department of Radiology, Duke University, Medical Center, Box 3302, Durham, NC 27710, USA*

Received 3 August 2001; accepted 30 October 2001

Abstract

Perceptrons are typically trained to minimize mean square error (MSE). In computer-aided diagnosis (CAD), model performance is usually evaluated according to other more clinically relevant measures. The purpose of this study was to investigate the relationship between MSE and the area (A_z) under the receiver operating characteristic (ROC) curve and the high-sensitivity partial ROC area ($_{0.90}A_z'$). A perceptron was used to predict lesion malignancy based on two mammographic findings and patient age. For each performance measure, the error surface in weight space was visualized. Comparison of the surfaces indicated that minimizing MSE tended to maximize A_z, but not $_{0.90}A_z'$. © 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Computer-aided diagnosis; Perceptron; Neural network; Breast cancer; Error surface

1. Introduction

While mammography is very sensitive at detecting breast cancer, its specificity is low. Only 15–34% of non-palpable, mammographically suspicious lesions are found to be malignant at biopsy [1,2]. The excessive number of benign breast biopsies raises the overall cost of mammographic screening to society [3] and results in emotional and physical burden to the patients. One goal of

☆ This work was supported in part by USPHS grant number R29-CA75547 awarded by the National Cancer Institute, Whitaker Foundation grant number RG 97-0322, Susan G. Komen Breast Cancer Foundation grant number 9803, and USAMRMC grants number DAMD 17-96-1-6226 and DAMD 17-94-J-4371 awarded by the US Army.

* Corresponding author. Department of Radiology, Duke University Medical Center, Box 3302, Durham, NC 27710, USA. Tel.: +1-919-684-7751; fax: +1-919-684-7122.
E-mail address: markey@duke.edu (M.K. Markey).

0010-4825/02/$ - see front matter
PII: S0010-4825(01)00035-X

the application of computer-aided diagnosis (CAD) to mammography is the reduction of this false positive rate.

In recent years, many breast cancer CAD studies have focused on the use of artificial neural network (ANN) models. ANN models have been developed to predict malignancy among suspicious breast lesions based upon mammographic and history findings [4–8]. Most networks for CAD are based on classic feed-forward, error-backpropagation paradigms, which are trained to minimize mean squared error (MSE) using a gradient descent technique. In "weight space", the ANN modifies a vector of weights, descending down a multi-dimensional error surface in search of the global minimum in MSE. Once trained, however, these ANNs are often evaluated according to other more clinically relevant measures of performance from receiver operating characteristic (ROC) analysis. Such measures include the ROC area index (A_z) and the partial area index ($_{0.90}A_z'$) corresponding to the portion of the ROC curve in the high sensitivity range of 0.9–1.0 [9,10]. (More information on the $_{0.90}A_z'$ is provided in the Methods section.)

The relationship between these three performance measures is not well defined, but there is a generally unstated assumption that a classifier trained to optimize MSE will also tend to optimize other measures such as A_z and $_{0.90}A_z'$. The validity of that assumption was questioned in recent studies. In one study, Kupinski et al. compared the performance of neural network models trained in the conventional manner (i.e., minimize MSE) vs. those trained by a niched Pareto multi-objective genetic algorithm (NP-GA) which simultaneously maximized sensitivity and specificity [11]. Using simulated XOR (exclusive or) data, they found that the ROC curve generated by NP-GA training was superior to that resulting from conventional training for both a perceptron (logistic discriminant) and an artificial neural network. Kupinski et al. also compared the performance of a conventionally trained perceptron to a NP-GA trained perceptron for the task of breast mass detection [12]. They found that while there was no significant difference between the models in terms of A_z, the NP-GA trained perceptron was significantly better in terms of the $_{0.90}A_z'$. In other words, the weights identified by minimizing the MSE were inferior to those identified by the NP-GA in terms of the model's performance at high sensitivities.

A related study demonstrated that different feature selection techniques might be preferred when $_{0.90}A_z$ is considered instead of A_z. Sahiner et al. compared the performance of linear discriminant analysis (LDA) classifiers using features selected by an LDA technique vs. a genetic algorithm (GA) [13]. The former provided better A_z but the latter had better $_{0.90}A_z'$.

All of the above studies examined the behavior of either linear or logistic discriminants. Although highly simplified compared to ANNs, these techniques are important for several reasons. First, their simplicity allows easy analysis of the relatively few parameters. For example, previous work at this institution presented a typical ANN for breast cancer CAD with 16 inputs and 10 hidden nodes, characterized by 180 weight parameters [14]. In comparison, the highly simplified perceptrons in this study were characterized by only four weights.

Secondly, several authors have reviewed recent studies where ANNs were applied to CAD problems, and suggested that a logistic model (such as a perceptron) would have likely provided similar performance while avoiding over-fitting problems [15,16]. Indeed, many recent studies in the field of CAD have been based upon linear discriminant models [17–20]. Any lessons learned from optimizing perceptrons would thus likely be useful to the field of CAD research.

The simple architecture of perceptrons is crucial to this study, which investigates the underlying behavior of these models by studying the error surfaces formed as a function of the parametric

weights. In particular, the goal is to compare error surfaces resulting from measuring performance with MSE vs. A_z and ${}_{0.90}A_z'$.

2. Materials and methods

2.1. Data set

The data set consisted of 500 cases of non-palpable breast lesions from patients who had undergone excisional biopsy at Duke University Medical Center between 1991 and 1996. In other words, the data set consisted of a consecutive sample of actual clinical cases. Of these 500 lesions, 65% were found to be benign as a result of histopathologic diagnosis. The relatively low prevalence of disease in this data set is consistent with the literature concerning this diagnostic task [1,2]. It is expected that models built on a clinically representative case mix will be better prepared to classify previously unseen clinical cases. The method of encoding the lesion descriptors has been previously described [14], and will only be summarized here. Expert radiologists retrospectively reviewed the patient films and recorded ten mammographic findings according to the Breast Imaging and Reporting Data System (BI-RADS™) lexicon [21], as well as other patient history data including the age. These findings were encoded into numeric values and used as input features in order to predict the known biopsy outcome of benign vs. malignant.

2.2. Network architecture

Even with the simplified architecture of a perceptron, it was still important to reduce the dimensionality of the input features in order to permit visualization and analysis. The number of inputs was therefore pruned to the three most important ones, based upon previous work in identifying the most important input findings for this diagnostic problem [14,22]. The BI-RADS™ findings used were mass margin and calcification morphology. In addition, a single patient history variable, age, was used. All features were scaled to the range of 0–1. This 3-input perceptron is shown in Fig. 1. The perceptron had one weight per input (W1, W2, and W3) and a bias term (W4). The dot product of input vector and the weight vector is passed through a non-linear activation function to produce the output. The inputs were the two BI-RADS™ findings, calcification morphology (weight W1) and mass margin (weight W2), and patient age (weight W3). The outputs of the perceptron range from 0, which indicates a benign lesion, to 1, which indicates a malignant lesion. Perceptron learning parameters were empirically optimized to minimize MSE: learning rate and momentum of 0.05 and 1000 iterations, with each iteration defined as a complete presentation of all training cases with weight adjustment after each case.

2.3. Error surface analysis

In weight space, each weight defines a dimension. Each point in the four-dimensional weight space represents a vector of weight values that define a distinct perceptron. When this perceptron is applied to a data set of input cases, the resulting MSE or other measures of performance are functions of the weights defining that perceptron. The error surface is the surface formed by evaluating the MSE

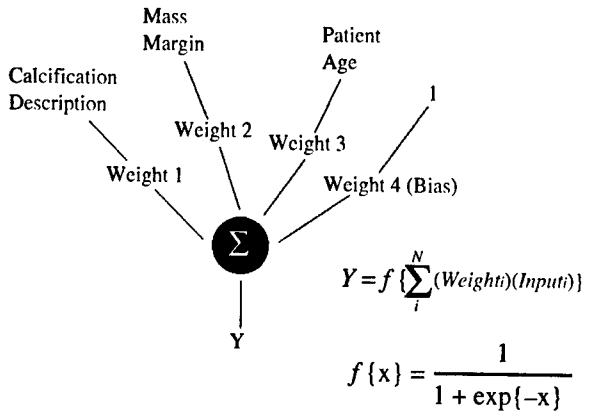

Fig. 1. Architecture of the perceptron. The dot product of the input vector (calcification morphology, mass margin, age, and bias) and the weight vector (Weight 1, Weight 2, Weight 3, and Weight 4) is passed through a non-linear activation function ($f(x)$) to produce the output (Y).

as a function of a range of weight values in each dimension. Other measures of performance, such as A_z and ${}_{0.90}A_z'$, can be used to form different surfaces in the same manner as the error surface. For simplicity, we refer to all such surfaces of performance measures as "error surfaces". Notice that plotting the error surface is not an optimization technique, but instead is used to show general trends in the data. For a perceptron with only two weights, the error surface may be readily plotted in the "z" or third dimension. In the current study, however, two-dimensional slices of the error surface are plotted instead of attempting to visualize the four-dimensional error surface. In a slice, two of the weights are varied to produce the surface, while the other two weights are held constant. Fig. 2 shows an example of an error surface slice. For simplicity, in the remainder of the error surface plots, the performance function will be plotted as intensity as in Fig. 3A.

To generate these slices, a grid search through weight space was performed. The perceptron with each combination of weights was applied to the data set. The MSE, ROC area (A_z), or partial area index (${}_{0.90}A_z'$) of each perceptron is indicated by intensity. Although the MSE and ROC have been reported in many previous studies, the ${}_{0.90}A_z'$ is relatively less well studied. ${}_{0.90}A_z'$ can be interpreted as the mean specificity of the model over the given high sensitivity range. It has particular clinical relevance in these examples of breast cancer CAD, where it is much more important to optimize sensitivity in the uppermost portion of the ROC curve, rather than specificity in the leftmost portion of the ROC curve. Note that while lower values for MSE indicate better performance, higher values for the performance measures A_z and ${}_{0.90}A_z'$ indicate better performance.

The ${}_{\mathrm{TPF_0}}A_z'$ was defined by Jiang et al. [10]. The partial area is the area under the ROC curve from a given sensitivity ($\mathrm{TPF_0}$) to 1.0, where $\mathrm{TPF_0} = 0.90$ is typically used. The partial area index (${}_{\mathrm{TPF_0}}A_z'$) is the partial area normalized by dividing by the constant ($1 - \mathrm{TPF_0}$). Note that the optimal value of both A_z and ${}_{0.90}A_z'$ is 1.0, but the chance behavior is 0.5 for A_z while it is 0.05 for ${}_{0.90}A_z'$ at $\mathrm{TPF_0} = 0.90$. The ROC analysis was performed using software modified and provided by Charles Metz, University of Chicago. The A_z and ${}_{0.90}A_z'$ were calculated using a modified version of the LABROC4 software, which finds a maximum likelihood estimate of the area from a fit to the data. The statistical comparisons were calculated using a modified version of the CLABROC software,

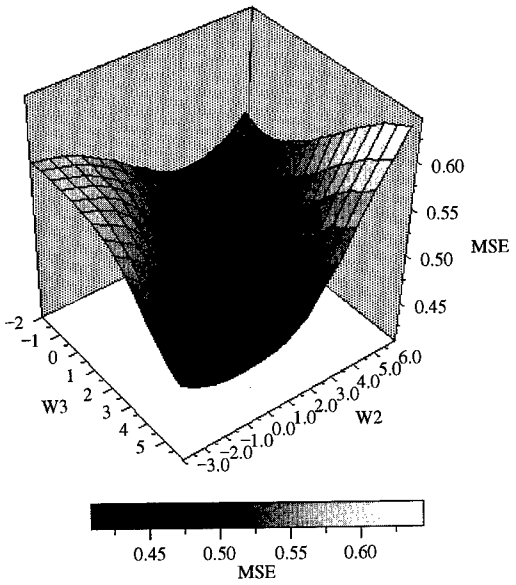

Fig. 2. A MSE surface in weight space. The MSE is a function of the perceptron weights (W1, W2, W3, and W4). W1 and W4 were held constant.

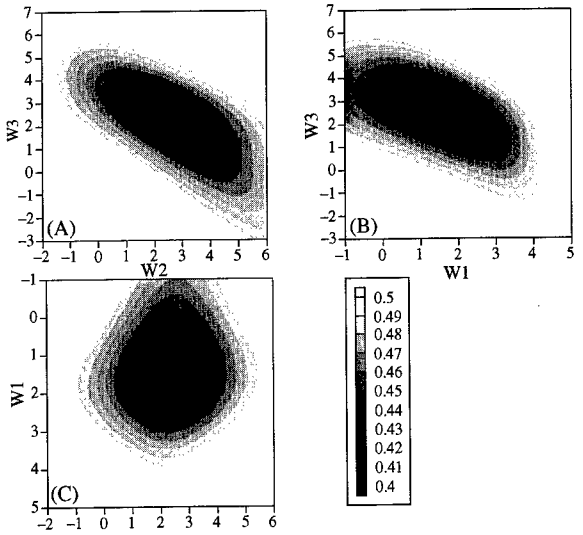

Fig. 3. The MSE surface in weight space. The MSE is a function of the perceptron weights (W1, W2, W3, and W4). The MSE is shown as intensity. Darker gray indicates better performance. The slices through MSE surface are (A) W3 vs. W2, (B) W3 vs. W1, and (C) W1 vs. W2. The subplots are arranged such that folding them into a box provides a way to visualize three of the weight dimensions.

which finds a maximum likelihood estimate of the areas for two classifications from fits to the two data sets. An estimate of statistical significance is reported for differences between the fitted curves. This estimate of significance includes the contribution from correlation of the input data.

The grid search over the weights was done in the vicinity of weights identified as optimal by training a perceptron to minimize the MSE of the data set. In other words, the training was used only to narrow down the reasonable range of weights over which the grid search was performed. With learning rate and momentum of 0.05 and 1000 iterations, the final weights were W1 = 1.65, W2 = 2.22, W3 = 2.56, and W4 = −3.21. In order to simplify the visualization further, the bias weight W4 was always fixed at that 'central' value. Each two-dimensional slice was generated by varying two of the feature weights while the bias and one remaining feature weight were held constant at the aforementioned 'central' values. The three combinations resulted in an "exploded box" showing the three-dimensional relationship between the three weights W1, W2, and W3. Each weight was varied approximately over the range of the central value ±150% of the central value. W1 was varied from −1.00 to 5.00. W2 was varied from −2.00 to 5.95. W3 was varied from −3.00 to 6.90.

3. Results

3.1. MSE vs. A_z

Fig. 3 shows three two-dimensional slices through the MSE surface and Fig. 4 shows three two-dimensional slices through the A_z surface. Note that improved performance corresponds to minimizing MSE (darker grayscale value) but maximizing A_z (brighter grayscale value). MSE is expected to range between 0 (perfect) and 0.5 (chance behavior), while A_z ranges between 0.5 (chance) and 1 (perfect). While the MSE and A_z surfaces are clearly not the same, the minimum observed on the MSE surface is in the same general location in weight space as the maximum

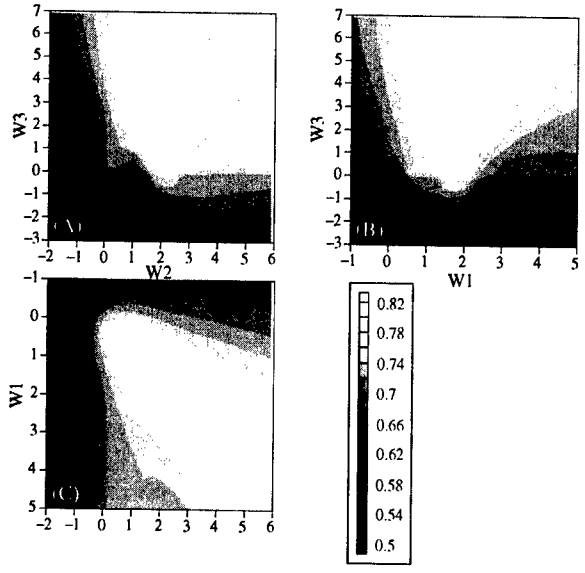

Fig. 4. The A_z surface in weight space. The A_z is a function of the perceptron weights (W1, W2, W3, and W4). The A_z is shown as intensity. Lighter gray indicates better performance. The slices through the A_z surface are (A) W3 vs. W2, (B) W3 vs. W1, and (C) W1 vs. W2.

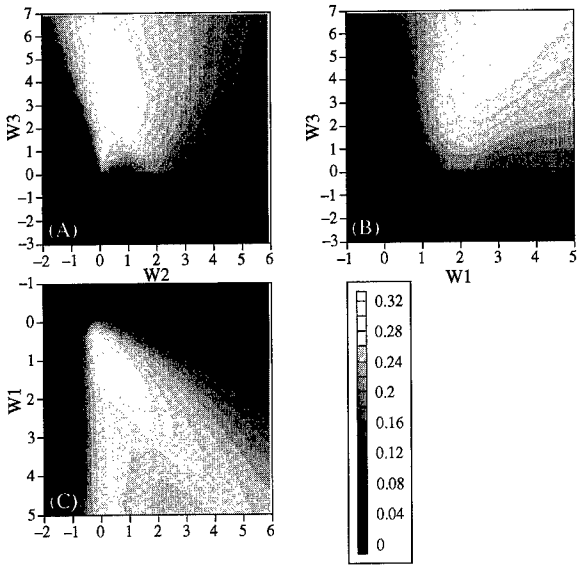

Fig. 5. The $_{0.90}A_z'$ surface in weight space. The $_{0.90}A_z'$ is a function of the perceptron weights (W1, W2, W3, and W4). The $_{0.90}A_z'$ is shown as intensity. Lighter gray indicates better performance. The slices through the $_{0.90}A_z'$ surface are (A) W3 vs. W2, (B) W3 vs. W1, and (C) W1 vs. W2.

observed on the A_z surface. The best solution corresponding to the global minimum on the MSE surface, i.e. the central weights (W1 = 1.65, W2 = 2.22, W3 = 2.56, and W4 = −3.21), has MSE of 0.41 and A_z of 0.80 ± 0.02. The best solution corresponding to the global maximum on the A_z surface (W1 = 1.65, W2 = 1.90, W3 = 2.40, W4 = −3.21, Fig. 4A) has MSE of 0.41 and A_z of 0.80 ± 0.02. The difference in the A_z between the solutions was not statistically significant (two tail $p = 0.14$).

3.2. MSE vs. $_{0.90}A_z'$

Fig. 3 shows three two-dimensional slices through the MSE surface and Fig. 5 shows three two-dimensional slices through the $_{0.90}A_z'$ surface. There is less correspondence in the general appearance of the contours between the MSE and $_{0.90}A_z'$ surfaces than was observed between MSE and A_z surfaces. The solution on the MSE surface, i.e. the central weights (W1=1.65, W2=2.22, W3=2.56, and W4 = −3.21) does not correspond to the best solution corresponding to a global maximum in the $_{0.90}A_z'$ surface (W1 = 3.35, W2 = 2.22, W3 = 5.70, and W4 = −3.21, Fig. 5B). The solution on the MSE surface has MSE of 0.41 and $_{0.90}A_z'$ of 0.24 ± 0.05. The solution on the $_{0.90}A_z'$ surface has MSE of 0.58 and $_{0.90}A_z'$ of 0.30 ± 0.04. The difference in $_{0.90}A_z'$ between the solutions was statistically significant (two tail $p = 0.006$).

This same trend may be demonstrated by comparing a particular operating point, such as the specificity for 95% sensitivity. The best MSE solution resulted in a specificity of 25% while the best specificity solution resulted in a specificity of 31%. This difference in specificity at 95% sensitivity was again statistically significant ($p = 0.002$).

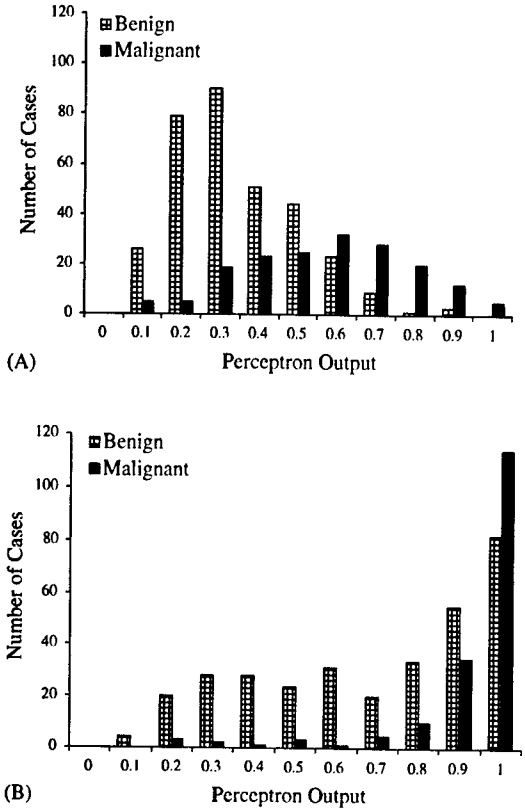

Fig. 6. Histograms of the outputs of the perceptron for the weights that correspond to (A) the minimal MSE and (B) the maximal $_{0.90}A_z'$.

The difference in the solutions on the MSE and $_{0.90}A_z'$ surfaces is illustrated by comparing the histograms of the outputs of the corresponding perceptrons (Fig. 6). Since the $_{0.90}A_z'$ measure describes the high sensitivity region of the ROC curve, the outputs of the perceptron with the highest $_{0.90}A_z'$ tend to be higher than the outputs of the perceptron with the lowest MSE.

4. Discussion

The three metrics of performance studied here are important for different reasons. The MSE is the metric that many models including perceptrons and ANNs attempt to optimize directly, while the A_z and $_{0.90}A_z'$ have greater clinical significance. Consider the histograms (Fig. 6) of network outputs of benign cases and malignant cases, where the network output of "0" indicates a benign lesion and "1" indicates a malignant lesion. MSE is a measure of the how close the distribution of benign cases is to a network output of "0" and how close the distribution of malignant cases is to "1". The area under the ROC curve is a measure of the overlap of the distributions. A training scheme that minimizes MSE, and so pulls the distributions to the edges, can also reduce the overlap of the distribution, and so increases A_z. It should be noted, however, that the MSE can decrease

without an accompanying change in A_z, because each increment in A_z can only result from the reversal of position for an adjacent pair of benign and malignant cases in the histogram. While a full convergence to MSE = 0 will also result in $A_z = 1$, the latter can be achieved with any arbitrary MSE, as long as the two distributions do not overlap at all. In the current study, it was observed that the weights that minimized MSE also maximized A_z.

It should be noted that in this study an ROC curve was generated by applying a threshold to the output node of the perceptron. By comparison, the method of Woods and Bowyer [23] scales the bias weight for the nodes in the hidden layer of an artificial neural network. Since perceptrons lack a hidden layer, their method would not be appropriate here.

In recent years, the sensitivity of breast cancer CAD techniques has been particularly emphasized, since there is a considerably greater cost in missing or delaying the diagnosis of an actual cancer (false negative) compared to referring a benign lesion to an unnecessary biopsy (false positive). For a range of sensitivities (e.g., TPF_0 from 0.9 to 1), the ${}_{TPF_0}A_z'$ can be thought of as an average specificity [10]. As an aid to interpreting these surfaces, it is helpful to note that for low values of the threshold TPF_0, the ${}_{TPF_0}A_z'$ surface resembles the A_z surface. Conversely, as TPF_0 increases, the ${}_{TPF_0}A_z'$ surface resembles the specificity surface at a given high sensitivity level. Unlike MSE and A_z, ${}_{0.90}A_z'$ is not symmetric in the sense that false negative and false positive cases do not contribute to the measure in the same way. In this work, the solution on the ${}_{0.90}A_z'$ surface was found to not correspond well with the MSE solution. It should be noted that the differences in the weights that optimize MSE vs. ${}_{0.90}A_z'$ may be due in part to biases inherent to the reduced amount of data that is associated with the high sensitivity region of the ROC curve.

If it is thought that A_z is a suitable measure of performance of CAD systems for breast cancer, then this work can be interpreted as a reassurance that classifiers trained to minimize MSE may also maximize the measure of interest. This provides some justification for avoiding the task of attempting to directly optimize model performance according to A_z. Note that optimizing for A_z by gradient descent techniques is not straightforward since A_z is not a continuous function.

However, if ${}_{0.90}A_z'$ corresponding to a given high level of sensitivity is a better measure of the quality of CAD systems for breast cancer, then this work demonstrates that a classifier trained to minimize MSE may provide an inferior solution. Alternative methods of identifying good weights for a perceptron or multi-layer network should be considered, such as evolutionary computing techniques that employ stochastic optimization. Our conclusions are consistent with related previous work that compared optimization techniques. As described in the introduction, Kupinski et al. found that using a perceptron (logistic discriminant) trained by a genetic algorithm instead of a classically trained perceptron resulted in no significant change in A_z, but a significant improvement in ${}_{0.90}A_z'$ [12].

5. Summary

Perceptrons, like more complicated backpropagation artificial neural networks, are typically trained to minimize mean square error (MSE). In computer-aided diagnosis (CAD) applications, model performance is usually evaluated according to other more clinically relevant measures from receiver operating characteristic (ROC) analysis. The purpose of this study was to investigate the relationship between MSE and the area (A_z) under the ROC curve and the partial ROC area (${}_{0.90}A_z'$) under the high sensitivity portion of the ROC curve. A perceptron was used to predict whether or not breast

lesions were malignant based on two mammographic findings and patient age. For each performance measure, the error surface in weight space was visualized. Comparison of the surfaces indicated that minimizing MSE tended to maximize A_z, but not ${}_{0.90}A_z'$. If it is important to maximize ${}_{0.90}A_z'$, then predictive models trained to minimize MSE may provide inferior solutions.

References

[1] D.B. Kopans, The positive predictive value of mammography, Am. J. Roentgenol. 158 (1992) 521–526.

[2] A.M. Knutzen, J.J. Gisvold, Likelihood of malignant disease for various categories of mammographically detected, nonpalpable breast lesions, Mayo Clin. Proc. 68 (1993) 454–460.

[3] D. Cyrlak, Induced costs of low-cost screening mammography, Radiology 168 (1988) 661–663.

[4] Y. Wu, M.L. Giger, K. Doi, C.J. Vyborny, R.A. Schmidt, C.E. Metz, Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer, Radiology 187 (1993) 81–87.

[5] C.E. Floyd Jr., J.Y. Lo, A.J. Yun, D.C. Sullivan, P.J. Kornguth, Prediction of breast cancer malignancy using an artificial neural network, Cancer 74 (1994) 2944–2948.

[6] J.A. Baker, P.J. Kornguth, J.Y. Lo, M.E. Williford, C.E. Floyd Jr., Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon, Radiology 196 (1995) 817–822.

[7] Y. Jiang, R.M. Nishikawa, D.E. Wolverton, C.E. Metz, M.L. Giger, R.A. Schmidt, C.J. Vyborny, K. Doi, Malignant and benign clustered microcalcifications: automated feature analysis and classification, Radiology 198 (1996) 671–678.

[8] H.P. Chan, B. Sahiner, N. Petrick, M.A. Helvie, K.L. Lam, D.D. Adler, M.M. Goodsitt, Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network, Phys. Med. Biol. 42 (1997) 549–567.

[9] D.K. McClish, Analyzing a portion of the ROC curve, Med. Decision Making 9 (1989) 190–195.

[10] Y. Jiang, C.E. Metz, R.M. Nishikawa, A receiver operating characteristic partial area index for highly sensitive diagnostic tests, Radiology 201 (1996) 745–750.

[11] M.A. Kupinski, M.A. Anastasio, Multiobjective genetic optimization of diagnostic classifiers with implications for generating receiver operating characteristic curves, IEEE Trans. Med. Imag. 18 (1999) 675–685.

[12] M.A. Kupinski, M.A. Anastasio, M.L. Giger, Multiobjective genetic optimization of diagnostic classifiers used in computerized detection of mass lesions in mammography, Presented at SPIE Medical Imaging 2000: Image Processing, 2000.

[13] B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, M.M. Goodsitt, Design of a high-sensitivity classifier based on a genetic algorithm: application to computer-aided diagnosis, Phys. Med. Biol. 43 (1998) 2853–2871.

[14] J.Y. Lo, J.A. Baker, P.J. Kornguth, C.E. Floyd Jr., Effect of patient history data on the prediction of breast cancer from mammographic findings with artificial neural networks, Acad. Radiol. 6 (1999) 10–15.

[15] G. Schwarzer, W. Vach, M. Schumacher, On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology, Stat. Med. 19 (2000) 541–561.

[16] D.J. Sargent, Comparison of artificial neural networks with other statistical approaches—results from medical data sets, Cancer 91 (2001) 1636–1642.

[17] M.L. Giger, H. Al-Hallaq, Z. Huo, C. Moran, D.E. Wolverton, C.W. Chan, W. Zhong, Computerized analysis of lesions in US images of the breast, Acad. Radiol. 6 (1999) 665–674.

[18] B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, M.M. Goodsitt, Computerized characterization of masses on mammograms: the rubber band straightening transform and texture analysis, Med. Phys. 25 (1998) 516–526.

[19] B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, L.M. Hadjiiski, Improvement of mammographic mass characterization using spiculation measures and morphological features, Med. Phys. 28 (2001) 1455–1465.

[20] M.F. McNitt-Gray, E.M. Hart, N. Wyckoff, J.W. Sayre, J.G. Goldin, D.R. Aberle, A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution CT: preliminary results, Med. Phys. 26 (1999) 880–888.

[21] BI-RADS, American College of Radiology Breast Imaging—Reporting and Data System (BI-RADS), 3rd Edition, American College of Radiology, Reston, VA, 1993–1998.

[22] J.Y. Lo, J.A. Baker, P.J. Kornguth, C.E. Floyd Jr., Computer-aided diagnosis of breast cancer: artificial neural network approach for optimized merging of mammographic features, Acad. Radiol. 2 (1995) 841–850.
[23] K. Woods, K.W. Bowyer, Generating ROC curves for artificial neural networks, IEEE Trans. Med. Imaging 16 (1997) 329–337.

Mia K. Markey is a doctoral candidate in Biomedical Engineering at Duke University. She received her B.S. from Carnegie Mellon University. She was awarded a Dissertation Research Award from the Susan G. Komen Breast Cancer Foundation. Her research interest is the application of machine learning techniques to problems in biology and medicine. Her current work is in computer-aided diagnosis of breast cancer, with a focus on the development of decision aids to reduce the number of benign breast biopsies.

Joseph Y. Lo received his B.S.E. in 1988 and Ph.D. in 1993, both in Biomedical Engineering from the Duke University School of Engineering. He is now an Assistant Research Professor in the Department of Radiology at the Duke University Medical Center and the Department of Biomedical Engineering at Duke University School of Engineering. His research interests involve image processing and artificial intelligence techniques applied to medical imaging, in particular the computer-aided diagnosis of breast cancer.

Rene Vargas-Voracek received his B.S. degree in Electrical Engineering from the National University of Mexico, Mexico City, in 1989 and the M.S. and Ph.D. degrees in Electrical Engineering from Duke University, Durham North Carolina in 1991 and 1996, respectively. He is currently a Research Associate at the Division of Imaging, Department of Radiology at Duke University Medical Center. His research interests include medical image and signal processing, information theory and statistical pattern classification and estimation.

Georgia D. Tourassi, Ph.D. is an Assistant Research Professor in the Department of Radiology at Duke University Medical Center. She earned a B.S. in Physics from the University of Thessaloniki in Greece and the Ph.D. in Biomedical Engineering from Duke University. Dr. Tourassi is a member of the Institute of Electrical and Electronics Engineers (IEEE), the International Society for Optical Engineering (SPIE), the American Association of Physicists in Medicine (AAPM), and the Radiological Society of North America (RSNA). She is also associate editor in *Radiology*. Her research interests include applications of artificial intelligence in computer-aided medical diagnosis and medical image processing.

Carey E. Floyd Jr. graduated from Eckerd College in 1976, earned the Ph.D. in Physics from Duke University in 1981, and is currently Professor of Radiology and Biomedical Engineering at Duke where he directs the Digital Imaging Research Division of Radiology. His research interests include computer assisted medical diagnosis and digital radiographic imaging.

Mia K. Markey, BS
Joseph Y. Lo, PhD
Carey E. Floyd, Jr, PhD

Index terms:
Breast neoplasms, 00.31, 00.32
Breast neoplasms, calcification, 00.81
Breast neoplasms, diagnosis, 00.129
Computers, diagnostic aid
Computers, neural network

Published online before print
10.1148/radiol.2232011257
Radiology 2002; 223:489–493

Abbreviations:
A_z = area under ROC curve
BI-RADS = Breast Imaging Reporting and Data System
BP-ANN = back-propagation artificial neural network
CAD = computer-aided diagnosis
LDA = linear discriminant analysis
PPV = positive predictive value
ROC = receiver operating characteristic

[1] From the Departments of Biomedical Engineering and Radiology, Digital Imaging Research Division, Duke University Medical Center, DUMC 3302, Durham, NC 27710. Received July 23, 2001; revision requested September 4; revision received October 12; accepted December 10. Supported in part by U.S. Public Health Service grants R29-CA75547, R21-CA092573, and R21-CA81309 awarded by the National Cancer Institute; Whitaker Foundation grants RG-97-0322 and SO-97-0035; U.S. Army Medical Research and Materiel Command grant DAMD17-99-1-9174 awarded by the U.S. Army; and Susan G. Komen Breast Cancer Foundation grants 9803 and BCTR2000730A. **Address correspondence to M.K.M.** (e-mail: *markey@duke.edu*).

Author contributions:
Guarantor of integrity of entire study, M.K.M.; study concepts and design, M.K.M., J.Y.L., C.E.F.; literature research, M.K.M., J.Y.L.; experimental studies, M.K.M., J.Y.L., C.E.F.; data acquisition and analysis/interpretation, M.K.M., J.Y.L., C.E.F.; statistical analysis, M.K.M., J.Y.L., C.E.F.; manuscript preparation, definition of intellectual content, editing, revision/review, and final version approval, M.K.M., J.Y.L., C.E.F.

Differences between Computer-aided Diagnosis of Breast Masses and That of Calcifications[1]

PURPOSE: To compare the performance of a computer-aided diagnosis (CAD) system for diagnosis of previously detected lesions, based on radiologist-extracted findings on masses and calcifications.

MATERIALS AND METHODS: A feed-forward, back-propagation artificial neural network (BP-ANN) was trained in a round-robin (leave-one-out) manner to predict biopsy outcome from mammographic findings (according to the Breast Imaging Reporting and Data System) and patient age. The BP-ANN was trained by using a large (>1,000 cases) heterogeneous data set containing masses and microcalcifications. The performances of the BP-ANN on masses and microcalcifications were compared with use of receiver operating characteristic analysis and a z test for uncorrelated samples.

RESULTS: The BP-ANN performed significantly better on masses than microcalcifications in terms of both the area under the receiver operating characteristic curve and the partial receiver operating characteristic area index. A similar difference in performance was observed with a second model (linear discriminant analysis) and also with a second data set from a similar institution.

CONCLUSION: Masses and calcifications should be considered separately when evaluating CAD systems for breast cancer diagnosis.

Among American women, breast cancer is the most common cancer and is the second leading cause of cancer deaths (1). Women in the United States have about a 1 in 8 lifetime risk of developing invasive breast cancer (2,3). Mammographic screening has been shown to reduce the mortality of breast cancer by as much as 30% (4,5). However, mammography has a low positive predictive value (PPV). Approximately 35% or less of women who undergo biopsy for histopathologic diagnosis of breast cancer are found to have malignancies (6). One goal of the application of computer-aided diagnosis (CAD) to mammography is to reduce the false-positive rate. Avoiding benign biopsies spares women unnecessary discomfort, anxiety, and expense.

CAD of breast cancer is the application of computational techniques to the problem of interpreting breast images, usually mammograms (7–9). There are two major topics in breast cancer CAD: detection of mammographic lesions and diagnosis of cancer from identified lesions. In the detection task, the goal is to assist a radiologist in the identification, and often the localization, of lesion-containing regions of mammograms. In the diagnosis task, the goal is to assist a radiologist in determining whether an identified breast lesion is an indication of cancer. This study focused on the diagnosis of breast lesions that had already been identified by radiologists as suspicious enough to warrant biopsy. In other words, these cases are generally considered indeterminate and more challenging, and any reduction in the number of benign biopsies represents an improvement over the status quo, provided high sensitivity is maintained.

Most breast biopsy is performed on lesions that manifest mammographically as either a mass or a cluster of microcalcifications (10). CAD systems for detection generally perform better on calcifications than on masses, as shown in two review articles (8,11) and a recent

study from a commercial CAD vendor (12). CAD systems for diagnosis that are based on features automatically extracted from the images are typically designed for either masses or calcifications alone. We are unaware of any previous attempts to compare the performance on masses and calcifications within a single study. Given the differences in databases and techniques with CAD systems for diagnosis, direct comparison of the published performances on masses and calcifications is not possible. However, the authors of classification studies on masses (13,14) report performances that are better than those reported in studies on calcifications (15,16). CAD systems for diagnosis that are based on findings extracted by radiologists are often trained and evaluated over heterogeneous data sets including both masses and calcifications, and the performances on masses and calcifications are not reported separately (17–20). The purpose of our study was to compare the performance of a CAD system for diagnosis of already detected lesions, based on radiologist-extracted findings on masses and calcifications.

MATERIALS AND METHODS

Data

Original studies were performed in accordance with standard clinical indications. All data from human subjects were collected with approval from appropriate institutional review boards, which also waived the requirement for informed patient consent.

We collected data on 1,530 nonpalpable mammographically suspicious breast lesions on which biopsy (core or excisional) was performed from 1990 to 2000 at Duke University Medical Center. The data were collected over several discontinuous time periods, but were collected consecutively within each time period. Of the 1,530 cases, 61 were removed because it was not certain that they were nonpalpable. In addition, 16 cases were removed because the radiologist's assessment of the likelihood of malignancy was unavailable. Thus, the primary data consisted of 1,453 approximately consecutive, nonpalpable, mammographically suspicious breast lesions. Experienced mammographers summarized each case according to the Breast Imaging Reporting and Data System (BI-RADS) lexicon (21). Each of the cases was read by one of seven readers. The 475 cases collected from 1990 to 1996 were read retrospectively, and the 978 cases collected from 1996 to 2000 were read prospectively.

Of the 1,453 cases, 508 (35%) were found to be malignant at biopsy. For the purposes of this study, a case was considered a "mass case" if mass features were present and no values were missing for any of the mass or calcification features. Likewise, a case was considered a "calcification case" if calcification features were present, but no mass features were present, and no values were missing for any of the mass or calcification features. There were 615 cases with masses, including 65 cases with calcifications in addition to a mass. There were 622 cases with calcifications that did not have masses as well. The PPVs for the mass cases (223/615 = 36%) and the calcification cases (209/622 = 34%) were similar ($P = .65$, χ^2 test for independence; 95% CI for malignancy fraction = −0.027, 0.080). The remaining 216 cases consisted of cases with neither a mass nor calcifications (n = 132) and cases with incomplete descriptions of the mass or calcifications that were present (n = 84). A mass was considered incompletely described if there were missing values for some of the mass or calcification features. Likewise, a calcification was considered incompletely described if there were missing values for some of the calcification features. The cases without a mass or calcifications were described by other findings, such as architectural distortion. When the value was missing for a feature, it was encoded in the same manner as if the finding was not present. All 1,453 cases, including the 216 cases with neither a mass nor calcifications, were used in building the CAD models for diagnosis.

A second data set consisted of 1,000 consecutive mammographically suspicious breast lesions on which excisional biopsy was performed from 1990 to 1997 at the University of Pennsylvania Medical Center. Experienced mammographers summarized each case according to the BI-RADS lexicon (21). Each of the cases was read retrospectively by one of 11 readers. Of the 1,000 cases, 396 (40%) were found to be malignant at biopsy. There were 481 cases with masses, including 10 cases with calcifications in addition to a mass. There were 449 cases with calcifications that did not also have masses. The PPV observed for the masses (191/481 = 40%) was the same as that for the calcifications (178/449 = 40%). There were 70 other cases, most (n = 68) of which were cases with incompletely described masses or calcifications. All 1,000 cases, including the incompletely described ones, were used in training the CAD models for diagnosis.

Specifically, the BI-RADS features collected were mass margin, mass shape, mass density, mass size, calcification morphology, calcification distribution, and associated and special findings. Although not a part of the BI-RADS specification, the number of calcifications is routinely collected at both institutions and was also included. The number of calcifications was indicated as no calcifications present, fewer than five, five to 10, or more than 10 calcifications present. The location of the lesion was also included and was encoded as posterior, central, axillary tail, subareolar, lower inner quadrant, lower outer quadrant, upper inner quadrant, or upper outer quadrant.

In addition to the BI-RADS findings, patient age was collected. For the cases from Duke University Medical Center, the mean age was 56 years, with a range of 23–87 years. For the cases from the University of Pennsylvania Medical Center, the mean age was 55 years, with a range of 17–92 years. Age is known to be an important risk factor for breast cancer. Increasing age is associated with increasing risk of breast cancer; a 60-year-old white American woman has a 14-fold increase in her chances of developing breast cancer relative to a 30-year-old white American woman (5). In agreement with the epidemiologic data, some evidence exists that age is a particularly valuable input in our predictive models (22).

For the cases from Duke University Medical Center, the mammographers indicated on a scale of 1–5 their assessment of the likelihood of malignancy. These assessment data were not available for the cases collected at the University of Pennsylvania Medical Center. An assessment of 1 indicated benign findings; 2, likely benign findings; 3, indeterminate findings; 4, likely malignant findings; and 5, malignant findings. The mammographer's assessment of malignancy was collected at the same time as the BI-RADS descriptors. As mentioned, some of the cases were read retrospectively and some were read prospectively, and although several mammographers participated in the study, each case was read by a single mammographer. Notice that this assessment is not the same as the BI-RADS clinical assessment. Moreover, this assessment does not directly correspond to the clinical task of deciding whether a patient should be referred to biopsy or follow-up. Since all the cases in the data set were subjected to biopsy, the mammographers were by definition performing with 100% relative sensitivity and 0% relative specificity on this data set (PPV,

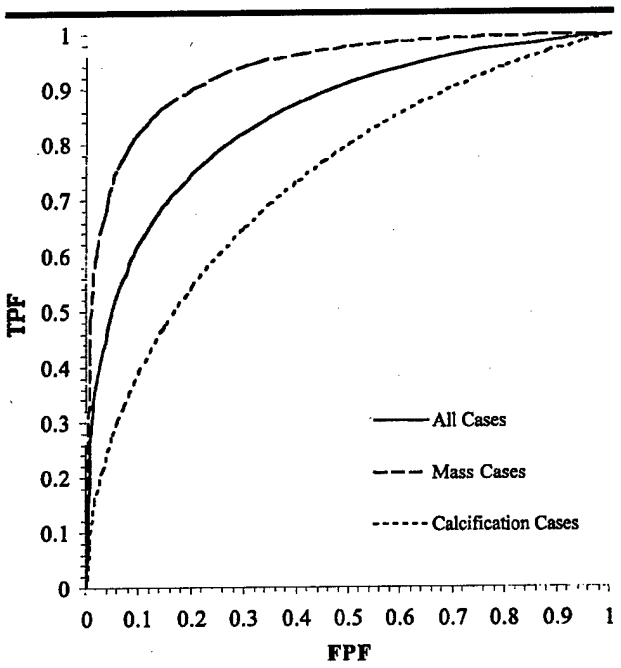

Figure 1. ROC curves for the mammographers' assessment of the likelihood of malignancy in the cases from Duke University Medical Center. The mammographers' assessment was more accurate for masses than for calcifications. *FPF* = false-positive fraction, *TPF* = true-positive fraction.

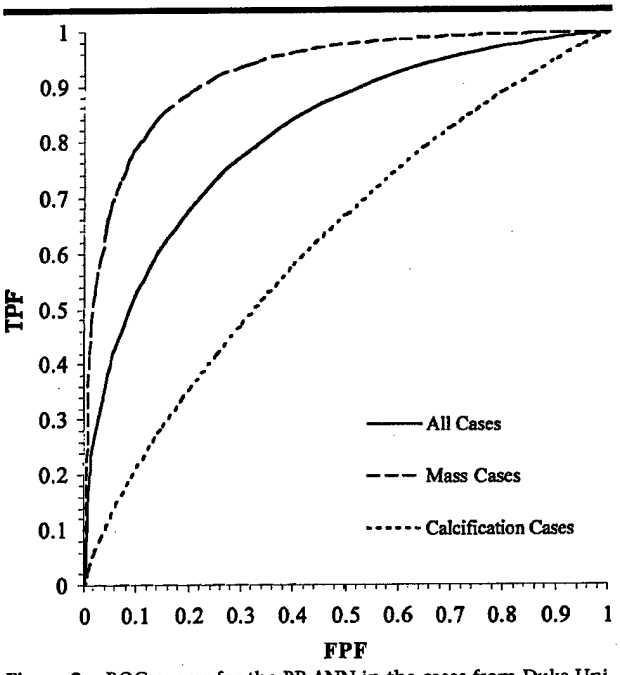

Figure 2. ROC curves for the BP-ANN in the cases from Duke University Medical Center. BP-ANN was more accurate for masses than for calcifications. *FPF* = false-positive fraction, *TPF* = true-positive fraction.

508/1,453 = 35%). (Notice that these relative measures are not indicative of the radiologists' performances over a general screening or diagnostic mammography patient population in which most actually benign cases are correctly referred to follow-up.) Nevertheless, their assessment of the likelihood of malignancy is useful as an approximation to an internal intermediate state in the decision process.

Artificial Neural Network

A feed-forward back-propagation artificial neural network (BP-ANN) can learn a function mapping inputs to outputs by being trained with cases of input-output pairs (23–25). The network inputs were the BI-RADS features and patient age. The network had a single hidden layer and one output node indicating malignancy. Each neuron in the network used a logistic activation function, $y = 1/(1 + e^{-x})$. The BP-ANN was trained to minimize the sum-of-squares error by using the back-propagation algorithm (23–25). A binary variable indicating benign or malignant was used as the network targets. The target values were clipped to 0.1 and 0.9 to ensure that the network weights remained finite (sigmoid units cannot produce 0 or 1). The network weights were updated after the presentation of each case (stochastic gradient descent), which can help alleviate the problem of local minima. A momentum term was used, which can also help the network escape local minima. The training cases were presented to the network in a round-robin (leave-one-out) manner. To avoid overtraining, network training ended when the average testing error on the left-out cases began to increase (early stopping). The network parameters (learning rate, momentum, and number of hidden nodes in the single hidden layer) were empirically optimized. The custom neural network software used was written by members of our laboratory and has been used in several previous publications (22).

Linear Discriminant Analysis

Linear discriminant analysis (LDA) was performed on the data collected at Duke University Medical Center. LDA is a common statistical technique for linear classification. The same input findings were used, and the cases were used in a round-robin fashion as with the BP-ANN. The LDA was computed by using the implementation in SAS software (SAS Institute, Cary, NC).

Receiver Operating Characteristic

The models were evaluated in terms of their receiver operating characteristic (ROC) curves. ROC curves enable the user to evaluate a model in terms of the trade-offs between sensitivity and specificity (26,27). The performance of classification methods can be evaluated by directly comparing their ROC curves or by comparing indices calculated from their curves. The most commonly used index is the area under the ROC curve (A_z). Notice that the values for A_z range from 0.5 for chance to 1.0 for a perfect classifier.

In breast cancer diagnosis, the decision task is whether to refer a suspicious case to biopsy or recommend follow-up imaging. A true-positive finding would be an actual cancer that was correctly referred to biopsy. A true-negative finding would be an actual benign lesion that was correctly recommended for follow-up imaging. The cost of missing a cancer (false-negative finding) far outweighs that of an unnecessary benign biopsy (false-positive finding). As a result, we were most concerned about the high sensitivity region of the curve, so we also used the partial area index ($_{0.90}A_z'$) calculated on that portion of the curve (true-positive fraction, 0.9–1.0) (28,29). The partial area index is the partial area normalized such that it ranges from 0.05 for chance to 1.0 for a perfect classifier. ROC analysis was performed by using software modified and

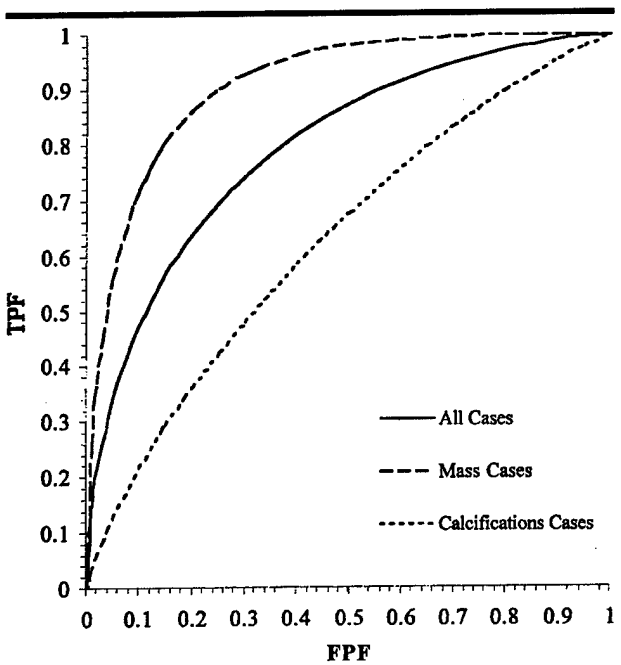

Figure 3. ROC curves for the LDA in the cases from Duke University Medical Center. LDA was more accurate for masses than for calcifications. *FPF* = false-positive fraction, *TPF* = true-positive fraction.

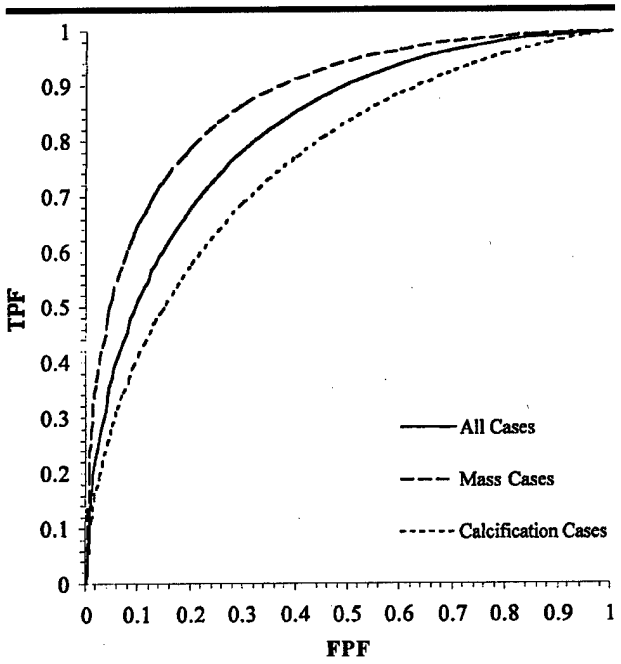

Figure 4. ROC curves for the BP-ANN in the cases from the University of Pennsylvania Medical Center. BP-ANN was more accurate for masses than for calcifications. *FPF* = false-positive fraction, *TPF* = true-positive fraction.

provided by Charles Metz at the University of Chicago. The modified LABROC4 software (maximum likelihood, semiparametric fit) was used to calculate the ROC curves and the curve indices, A_z and ${}_{0.90}A_z'$. Statistical comparisons were made with use of a standard z test since there was no correlation between the mass and calcification cases. A P value of less than .01 was considered to indicate a statistically significant difference.

RESULTS

Duke University Medical Center

Mammographers' assessment.—The mammographers' assessment of the likelihood of malignancy (five-point scale) was used as a decision variable, and ROC curves were formed for masses and calcifications separately (Fig 1). There was a significant difference ($P < .01$) in the ROC areas for the masses ($A_z = 0.94 \pm 0.01$) compared with that for the calcifications ($A_z = 0.74 \pm 0.02$). There was also a significant difference ($P < .01$) in the partial area index for the masses (${}_{0.90}A_z' = 0.62 \pm 0.06$) versus that for the calcifications (${}_{0.90}A_z' = 0.17 \pm 0.04$). The ROC curve over all of the cases was intermediate ($A_z = 0.85 \pm 0.01$, ${}_{0.90}A_z' = 0.34 \pm 0.04$). The assessment of the mammographers was more accurate for the masses than for the calcifications. Notice, however, that the actual clinical performance of the mammographers was essentially the same for masses (PPV = 223/615 = 36%) and calcifications (PPV = 209/622 = 34%, $P = .65$, χ^2 test for independence; 95% CI for malignancy fraction = −0.027, 0.080). Notice as well that since each case was read by a single mammographer and the study included seven readers, the assessment was pooled across mammographers.

BP-ANN performance.—The BP-ANN developed by using round-robin sampling on all of the cases from Duke University Medical Center also performed better on the masses than the calcifications (Fig 2). The difference in the ROC area for the masses ($A_z = 0.93 \pm 0.01$) and that for the calcifications ($A_z = 0.63 \pm 0.02$) was significant ($P < .01$). The difference in the partial area index was also significant ($P < .01$) between the masses (${}_{0.90}A_z' = 0.62 \pm 0.05$) and the calcifications (${}_{0.90}A_z' = 0.10 \pm 0.02$). The ROC curve over all of the cases was intermediate ($A_z = 0.82 \pm 0.01$, ${}_{0.90}A_z' = 0.30 \pm 0.03$).

Linear discriminant analysis.—The round-robin LDA classifier on the cases from Duke University Medical Center also performed better on the masses than on the calcifications (Fig 3). There was a significant difference ($P < .01$) in the ROC area for the masses ($A_z = 0.91 \pm 0.01$) versus that for the calcifications ($A_z = 0.62 \pm 0.02$). The difference in the partial area index between the masses (${}_{0.90}A_z' = 0.61 \pm 0.04$) and that for the calcifications (${}_{0.90}A_z' = 0.11 \pm 0.02$) was also significant ($P < .01$). The ROC curve over all of the cases was intermediate ($A_z = 0.80 \pm 0.01$, ${}_{0.90}A_z' = 0.28 \pm 0.03$).

University of Pennsylvania Medical Center: BP-ANN

The BP-ANN developed by using round-robin sampling on the cases from the University of Pennsylvania Medical Center also performed better on the masses than on the calcifications (Fig 4). There was a significant difference ($P < .01$) in the ROC area of the masses ($A_z = 0.88 \pm 0.02$) compared with that for the calcifications ($A_z = 0.76 \pm 0.02$). There was also a significant difference ($P < .01$) in the partial area index of the masses (${}_{0.90}A_z' = 0.45 \pm 0.05$) versus the calcifications (${}_{0.90}A_z' = 0.23 \pm 0.04$). The ROC curve over all of the cases was intermediate ($A_z = 0.82 \pm 0.01$, ${}_{0.90}A_z' = 0.34 \pm 0.03$).

DISCUSSION

In this study, the performances of a breast cancer CAD model on mass and microcalcification lesions were com-

pared. BP-ANN and LDA models were considered. BP-ANN analysis was repeated with data from a second similar institution. The mammographers' assessment of malignancy was also investigated. The performance on masses was consistently better than the performance on calcifications in comparisons involving radiologists, CAD models, and data from two institutions.

A BP-ANN trained in a round-robin fashion on a heterogeneous set of biopsy-proved breast lesions was found to perform significantly better on masses than calcifications in terms of the ROC area and the partial area index. This difference was seen with use of two data sets collected at different institutions, which argues that this phenomenon is not a function of a particular data set. A similar difference in performance on masses and calcifications was seen when another predictive model, LAD, was used. Moreover, in a separate study conducted at Duke University Medical Center, a similar difference in performance was observed with a constraint satisfaction neural network (30). This indicates that the observed performance differential is not specific to BP-ANN models. However, it is possible that if some other classification technique were used, such differences would not be observed between masses and calcifications. Finally, when the mammographers' assessment of the likelihood of malignancy was used as a decision variable, it was found that they too seemed to be able to more accurately assess the masses than the calcifications. Notice, however, that there is no corresponding difference in their clinical recommendations, based on the PPV of biopsy for those two subsets of cases. Taken together, these findings suggest that masses and calcifications should be considered separately when evaluating CAD systems for breast cancer diagnosis. It should be recalled that the "masses" in this study included both calcified and noncalcified masses and that the presence of calcifications in addition to a primary mass lesion may affect the classification of that mass by either a computational technique or a mammographer.

Recent work by Huo et al (14,31) describes a CAD system for diagnosis of breast masses that handles spiculated and nonspiculated masses separately and is superior to a CAD system that was developed on a mixture of spiculated and nonspiculated masses. The work described herein can be interpreted as further evidence of the effect of distinct subsets on the performance of the breast cancer CAD models for diagnosis. As larger databases become available for developing CAD models for diagnosis, it may be beneficial to develop modular systems with submodels that are specialized for subsets of the data. Alternatively, when a single CAD model for diagnosis is developed over a heterogeneous data set, such as one containing both mass and calcification cases, these results suggest that it would be appropriate to evaluate the performance of the overall model over the subsets of interest.

Acknowledgments: The authors thank the members of the breast imaging sections at Duke University Medical Center and the University of Pennsylvania Medical Center. We also acknowledge Brian Harrawood, MS, for scientific programming.

References

1. Ries LAG, Wingo PA, Miller DS, et al. The annual report to the nation on the status of cancer, 1973–1997, with a special section on colorectal cancer. Cancer 2000; 88:2398–2424.
2. Feuer EJ, Wun L, Boring CC, Flanders WD, Timmel MJ, Tong T. The lifetime risk of developing breast cancer. J Natl Cancer Inst 1993; 85:892–897.
3. Wun L, Merrill RM, Feuer EJ. Estimating lifetime and age-conditional probabilities of developing cancer. Lifetime Data Anal 1998; 4:169–186.
4. Shapiro S. Screening: assessment of current studies. Cancer 1994; 74:231–238.
5. Henderson IC. Breast cancer. In: Murphy GP, Lawrence W Jr, Lenhard RE, eds. American Cancer Society textbook of clinical oncology. Atlanta, Ga: American Cancer Society, 1995; 198–219.
6. Kopans DB. The positive predictive value of mammography. AJR Am J Roentgenol 1992; 158:521–526.
7. Doi K, MacMahon H, Katsuragawa S, Nishikawa RM, Jiang Y. Computer-aided diagnosis in radiology: potential and pitfalls. Eur J Radiol 1999; 31:97–109.
8. Vyborny CJ, Giger ML, Nishikawa RM. Computer-aided detection and diagnosis of breast cancer. Radiol Clin North Am 2000; 38:725–740.
9. Giger ML. Computer-aided diagnosis of breast lesions in medical images. Comput Sci Eng 2000; 2:39–45.
10. Liberman L, Abramson AF, Squires FB, Glassman JR, Morris EA, Dershaw DD. The Breast Imaging Reporting and Data System: positive predictive value of mammographic features and final assessment categories. AJR Am J Roentgenol 1998; 171:35–40.
11. Karssemeijer N, Hendriks JH. Computer-assisted reading of mammograms. Eur Radiol 1997; 7:743–748.
12. Castellino RA, Roehrig J, Zhang W. Improved computer-aided detection (CAD) algorithms for screening mammography (abstr). Radiology 2000; 217(P):400.
13. Chan HP, Sahiner B, Helvie MA, et al. Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study. Radiology 1999; 212:817–827.
14. Huo Z, Giger ML, Vyborny CJ, Wolverton DE, Schmidt RA, Doi K. Automated computerized classification of malignant and benign masses on digitized mammograms. Acad Radiol 1998; 5:155–168.
15. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Giger ML, Doi K. Improving breast cancer diagnosis with computer-aided diagnosis. Acad Radiol 1999; 6:22–33.
16. Chan HP, Sahiner B, Lam KL, et al. Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces. Med Phys 1998; 25:2007–2019.
17. Wu Y, Giger ML, Doi K, Vyborny CJ, Schmidt RA, Metz CE. Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer. Radiology 1993; 187:81–87.
18. Baker JA, Kornguth PJ, Lo JY, Williford ME, Floyd CE Jr. Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon. Radiology 1995; 196: 817–822.
19. Kahn CE Jr, Roberts LM, Shaffer KA, Haddawy P. Construction of a Bayesian network for mammographic diagnosis of breast cancer. Comput Biol Med 1997; 27:19–29.
20. Floyd CE Jr, Lo JY, Tourassi GD. Case-based reasoning computer algorithm that uses mammographic findings for breast biopsy decisions. AJR Am J Roentgenol 2000; 175: 1347–1352.
21. American College of Radiology. BI-RADS: American College of Radiology Breast Imaging Reporting and Data System (BI-RADS). 3rd ed. Reston, Va: American College of Radiology, 1998.
22. Lo JY, Baker JA, Kornguth PJ, Floyd CE Jr. Effect of patient history data on the prediction of breast cancer from mammographic findings with artificial neural networks. Acad Radiol 1999; 6:10–15.
23. Rumelhart DE, McClelland JL, ed. Parallel distributed processing: explorations in the microstructures of cognition. Cambridge, Mass: MIT Press, 1986.
24. Bishop CM. Neural networks for pattern recognition. Oxford, England: Oxford University Press, 1995.
25. Hertz J, Anders K, Palmer RG. Introduction to the theory of computation: Santa Fe Institute Studies in the Science of Complexity. Redwood City, Calif: Addison-Wesley, 1991.
26. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978; 8:283–298.
27. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986; 21:720–733.
28. McClish DK. Analyzing a portion of the ROC curve. Med Decis Making 1989; 9:190–195.
29. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology 1996; 201:745–750.
30. Tourassi GD, Markey MK, Lo JY, Floyd CE Jr. A neural network approach to breast cancer diagnosis as a constraint satisfaction problem. Med Phys 2001; 28:804–811.
31. Huo Z, Giger ML, Metz CE. Effect of dominant features on neural network performance in the classification of mammographic lesions. Phys Med Biol 1999; 44: 2579–2595.

ELSEVIER Artificial Intelligence in Medicine 27 (2003) 113–127

Artificial Intelligence in Medicine

www.elsevier.com/locate/artmed

Self-organizing map for cluster analysis of a breast cancer database

Mia K. Markey[a,b,*], Joseph Y. Lo[a,b], Georgia D. Tourassi[b], Carey E. Floyd Jr.[a,b]

[a]*Department of Biomedical Engineering, Duke University, Durham, NC 27708, USA*
[b]*Digital Imaging Research Division, Department of Radiology, Duke University Medical Center, Durham, NC 27710, USA*

Received 10 May 2002; received in revised form 1 November 2002; accepted 10 December 2002

Abstract

The purpose of this study was to identify and characterize clusters in a heterogeneous breast cancer computer-aided diagnosis database. Identification of subgroups within the database could help elucidate clinical trends and facilitate future model building. A self-organizing map (SOM) was used to identify clusters in a large (2258 cases), heterogeneous computer-aided diagnosis database based on mammographic findings (BI-RADS™) and patient age. The resulting clusters were then characterized by their prototypes determined using a constraint satisfaction neural network (CSNN). The clusters showed logical separation of clinical subtypes such as architectural distortions, masses, and calcifications. Moreover, the broad categories of masses and calcifications were stratified into several clusters (seven for masses and three for calcifications). The percent of the cases that were malignant was notably different among the clusters (ranging from 6 to 83%). A feed-forward back-propagation artificial neural network (BP-ANN) was used to identify likely benign lesions that may be candidates for follow up rather than biopsy. The performance of the BP-ANN varied considerably across the clusters identified by the SOM. In particular, a cluster (#6) of mass cases (6% malignant) was identified that accounted for 79% of the recommendations for follow up that would have been made by the BP-ANN. A classification rule based on the profile of cluster #6 performed comparably to the BP-ANN, providing approximately 25% specificity at 98% sensitivity. This performance was demonstrated to generalize to a large (2177) set of cases held-out for model validation.

Keywords: Self-organizing map; Cluster analysis; Breast cancer; Computer-aided diagnosis

* Corresponding author. Present address: Department of Biomedical Engineering, The University of Texas at Austin, ENS617B C0800, Austin, TX, USA. Tel.: +1-512-471-1711; fax: +1-512-471-0616.
E-mail address: mia.markey@mail.utexas.edu (M.K. Markey).

0933-3657/03/$ – see front matter
doi:10.1016/S0933-3657(03)00003-4

1. Introduction

There is considerable interest in the use of computational techniques to aid in the detection and diagnosis of breast cancer [5,8,26]. Most computer-aided diagnosis (CAD) studies, including this one, focus on mammography since it is the primary tool for the detection of breast lesions and the subsequent decision to biopsy suspicious lesions. The decision to biopsy is complicated by the fact that breast cancer can present itself in a variety of ways on a mammogram and there is considerable overlap in the appearance of benign and malignant lesions. CAD systems for the decision to biopsy that are based on findings extracted by radiologists are often trained and evaluated over heterogeneous databases that reflect this variability in the morphological appearance of suspicious breast lesions [1,7,28]. We have recently shown that a CAD tool trained on such a heterogeneous database can perform very differently on two broad subgroups which constitute most of the currently biopsied lesions: masses and microcalcifications [17]. In particular, we observed that the performance was significantly better on masses than on calcifications.

In this study, we used a self-organizing map (SOM) [13] to identify clusters in a heterogeneous breast cancer CAD database. SOM is an unsupervised learning method that relates similar input vectors to the same region of a map of neurons. To the best of our knowledge, SOMs have not been used to identify clusters in a CAD database similar to the one presented here. SOMs have been used for other tasks in breast cancer CAD such as a benchmark for model selection [27] and to predict biopsy outcome [4].

Once the SOM was used to identify the clusters, a constraint-satisfaction neural network (CSNN) was used to characterize the clusters by determining a profile for each cluster. Briefly, the CSNN is a Hopfield-type network of neurons arranged in a non-hierarchical way (Fig. 1). There are symmetric, bi-directional weights between all pairs of neurons but there are no reflexive weights. The CSNN operates as a nonlinear, dynamic system that tries to reach a globally stable state by adjusting the activation levels of the neurons under the constraints imposed by the a priori fixed weight values. A cluster "profile" provides a description of a "typical" case in the cluster. We have previously introduced CSNN for predicting biopsy outcome and as a data mining tool for breast cancer CAD databases [25].

A feed-forward back-propagation artificial neural network (BP-ANN) is a classic technique that is commonly used in breast cancer CAD systems. Consequently, a BP-ANN was used to predict the biopsy outcome [2,10,21] and the performance of the BP-ANN was compared on the clusters identified by the SOM and profiled by the CSNN.

A clustering algorithm such as an SOM followed by a cluster characterization method such as CSNN profiling could serve as tools in the initial phases of a divide-and-conquer approach to the computer-aided diagnosis of breast cancer. Both modular and ensemble methods could be used for a divide-and-conquer approach. A modular system uses multiple classifiers to solve a classification problem by partitioning the input space into smaller domains, each of which is handled by a local model [24]. The local models can be thought of as experts for a particular kind of case. Ensemble methods are resampling schemes in which the same cases are used in training multiple experts, whose predictions are then combined [24]. Such approaches may be justified in light of recent results in this field. Simple ensembles of classifiers using voting or averaging to combine their predictions have shown promise in computer-aided detection of breast masses [14,22,31]. Zheng et al.

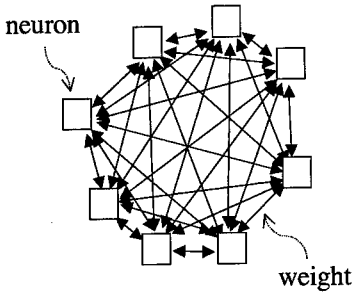

Fig. 1. Schematic of the constraint satisfaction neural network (CSNN). Notice that the neurons are fully interconnected with no reflexive weights.

employed a modular scheme, in which the data were partitioned by a difficulty measure, for computer-aided detection of breast masses with encouraging results [30]. Zheng et al. also investigated a promising ensemble of modular models, formed by taking the average of the predictions from modular models in which the data were partitioned using three features [29]. Huo and coworkers described a modular system, in which the data were partitioned by a spiculation measure, which was superior to a general image-based computer-aided diagnosis system [11,12]. Finally, we have recently demonstrated that a BI-RADS™-based CAD tool built on a heterogeneous database can perform very differently on two broad subgroups of lesions, masses and microcalcifications [17]; the CAD tools investigated performed better on masses than on calcifications. In all of the examples listed here, a priori knowledge was used to partition the data into subsets. Unsupervised learning may provide an alternate avenue to a priori knowledge for identifying subsets in the data that should be handled separately in the development or evaluation of computer-aided diagnosis or detection systems.

2. Materials and methods

2.1. Data

Approximately half of the available data (4435) were used for model development (2258) in this study in order to withhold the remaining data for additional model validation (2177); the data were randomly partitioned into the training and validation sets, but attention was paid to key summary statistics such as the fraction of cases that were malignant in each set. For each lesion, the benign or malignant status from pathologic diagnosis was known. The overall malignancy fraction was 43%. In the next few paragraphs, we describe the data (2258) used for model development in greater detail.

The first data set consisted of 751 non-palpable, mammographically suspicious breast lesions that underwent biopsy (core or excisional) at Duke University Medical Center from 1990 to 2000. The data collection procedures have been previously described [16]. Briefly, expert mammographers described each case using the breast imaging and reporting data system (BI-RADS™) lexicon [20]. Each of the cases was read by one of seven readers.

When a lesion could be described by multiple descriptors (e.g. pleomorphic and punctate), the mammographers were requested to report the descriptor that was most suspicious for malignancy (e.g. pleomorphic). Of the 751 cases, 260 (35%) were malignant.

The second data set consisted of 501 mammographically suspicious breast lesions that underwent excisional biopsy at the University of Pennsylvania Medical Center from 1990 to 1997. The data collection procedures have been previously described [16]. Briefly, each of the cases was read by one of 11 expert mammographers who described each case using the BI-RADS™ lexicon [20]. When a lesion could be described by multiple descriptors (e.g. pleomorphic and punctate), the mammographers were requested to report the descriptor that was most suspicious for malignancy (e.g. pleomorphic). Of the 501 cases, 200 (40%) were malignant.

The third data set consisted of 1006 biopsy-proven breast lesions randomly selected from the Digital Database for Screening Mammography [9]. Expert mammographers described each case using the BI-RADS™ lexicon [20]. Lesions that were described by multiple descriptors were encoded for our purposes using the descriptor that was most suspicious for malignancy. Of the 1006 cases, 522 (52%) were malignant.

Specifically, the six BI-RADS™ features collected describe the mass margin, mass shape, calcification morphology, calcification distribution, associated, and special findings. Missing values were encoded as zero. Each BI-RADS™ feature was encoded using uniformly scaled rank ordered categories (Table 1). For example, when a mass is present for a case, the mass margin can take on one of five values: well circumscribed (1), microlobulated (2), obscured (3), ill-defined (4), or spiculated (5). In addition to the BI-RADS™ features, the patient age was collected, for a total of seven features.

2.2. *Self-organizing map*

A self-organizing map relates similar cases (input vectors) to the same region of a map of neurons [13]. The SOM was computed using the SOM toolbox in MATLAB® (The MathWorks Inc., Natick, MA). The basic SOM consisted of 16 neurons arranged in a single layer in a 2-D square grid of 4×4 neurons, but different configurations were considered. For each case, the Euclidean distance between the case and each neuron was calculated based on the seven input features (the biopsy outcome was not provided to the SOM). For input to the SOM, each feature was scaled by subtracting the mean and dividing by the standard deviation, resulting in each scaled feature having mean zero and standard deviation of one. After the most similar neuron was determined the neurons in its neighborhood were identified. The neighborhood of a neuron was defined as all the neurons within a given link distance of the matched neuron. All the neurons in the neighborhood were adjusted to have feature values closer to the current case. The amount that the neuron weights were adjusted was controlled by the learning rate. The learning rates and distance threshold values used were the default values for the SOM toolbox.

2.3. *Constraint satisfaction neural network*

After the clusters were identified, a CSNN was used to determine the profiles of the clusters [23,25]. Custom software in the C language was used to implement the CSNN and

Table 1
Encoding of the BI-RADS™ features

Mass margin	Mass shape	Calcification morphology	Calcification distribution	Associated findings	Special findings
0: no mass	0: no mass	0: no calcifications	0: no calcifications	0: none	0: none
1: well circumscribed	1: round	1: milk of calcium like	1: diffuse	1: skin lesion	1: intrmam. lymph node
2: mircolobolated	2: oval	2: eggshell or rim	2: regional	2: hematoma	2: asymmetric breast tissue
3: obscured	3: lobulated	3: skin	3: segmental	3: post surgical scar	3: focal asymmetric density
4: ill-defined	4: irregular	4: vascular	4: linear	4: trabecular thickening	4: tubular density
5: spiculated		5: spherical or lucent centered	5: clustered	5: skin thickening	
		6: suture		6: skin retraction	
		7: coarse		7: nipple retraction	
		8: large rod-like		8: axillary adenopathy	
		9: round		9: architectural distortion	
		10: dystrophic			
		11: punctate			
		12: indistinct			
		13: pleomorphic			
		14: fine branching			

has been previously described [25]. The Lyapunov energy function was used as a measure of the network stability. It was found that 1000 iterations were sufficient to achieve stability. The weights were predetermined using autoassociative backpropagation neural networks (auto-BP). In keeping with our previous work [25], the auto-BP networks were trained with a learning rate of 1.0 for 100 iterations and the root mean squared training error was approximately 0.1 (network outputs between 0 and 1).

For each cluster, a CSNN was used to generate a profile. Each category of the categorical BI-RADS™ features corresponded to a binary variable and associated neuron. For example, the mass margin with its five non-zero categories was represented by five separate neurons. Patient age was translated into a discrete variable with five levels (<40 years, $40 \leq x < 50$, $50 \leq x < 60$, $60 \leq x < 70$, ≥ 70 years) [25]. An additional neuron was used to signify cluster membership. The activation level of the neuron indicating cluster membership was set to the maximal value and the other neurons were allowed to evolve until the network reached a stable state. The feature neurons that were activated defined the profile of the cluster. A profile is a list of feature values that succinctly summarizes the cluster and defines a "typical" case (e.g. mass margin is well circumscribed, mass shape is round, and patient age is between 50 and 60 years). All cases in the cluster do not exactly match the profile; there is still a distribution of feature values. Notice that unlike common summary statistics, such as the cluster centroid, the CSNN profile implicitly includes feature selection; only features deemed relevant to the network for describing a cluster are included.

2.4. Back-propagation artificial neural network (BP-ANN)

A feed-forward back-propagation artificial neural network (BP-ANN) was used to predict the biopsy outcome from the mammographic findings and patient age. The BP-ANN was trained to minimize the sum-of-squares error using the back-propagation algorithm [2,10,21]. The network had a single hidden layer of 14 neurons and each neuron in the network used a logistic activation function. The network inputs (7) were the BI-RADS™ features and patient age. Network inputs were rescaled from 0 to 1 (by subtracting the minimum value and dividing by the maximum minus the minimum). The biopsy outcomes were the network targets; there was one output node indicating malignancy. The 2258 cases were presented to the network in a round-robin manner (leave-one-out, k-fold cross-validation with $k = N$) and training ended before the average testing error on the left-out cases began to increase. The custom neural network software used was written in C++ by members of our laboratory, and the training and testing process has been reported previously [15,17].

2.5. Receiver operating characteristic

Receiver operating characteristic (ROC) curves can be used to show the trade-off in sensitivity and specificity achievable by a classifier by varying the threshold on the output decision variable [18,19]. The area under the ROC curve is often used as a measure of classifier performance. In evaluating models for diagnosing breast cancer, all sensitivities are not of equal interest. Only techniques that perform with very high sensitivity would be

clinically acceptable since missing a cancer (false negative) is generally considered much worse that an unnecessary benign biopsy (false positive). Thus, particular attention was paid to the specificity at 98% sensitivity.

The ROC curves were calculated non-parametrically. *P*-values and standard deviations on the specificity at 98% sensitivity were estimated by bootstrap sampling on the decision variable [6].

3. Results

Fig. 2 illustrates the arrangement of the neurons in the SOM. The set of cases that were mapped to a neuron defined a cluster. Fig. 2 shows the number of cases that were mapped to each neuron, i.e. the number of cases in each cluster. The fraction of the cases in each cluster that were malignant is also shown in Fig. 2 (bottom number in italics). The malignancy fraction is not shown for the clusters with fewer than 10 cases (#5, 12, and 15), on the assumption that no meaningful conclusions can be drawn from such a small number of cases. Inspection of the cases mapped to these clusters (#5, 12, and 15) revealed that the cases are rare for this database. They included cases with findings that were seen with a very low prevalence in the set (e.g. special finding of intramammary lymph node) or reflected incomplete or inconsistent data (e.g. the calcification morphology was described but calcification distribution feature was not reported). Together these three clusters comprise only 0.5% of the cases. Therefore, no further analysis was performed on these clusters. Recall that the SOM was not provided with the biopsy outcome information. The differences in the malignancy fraction are a reflection of differences in the BI-RADS™ features and patient age between the clusters. Cluster malignancy rates near 50% do contain some information since the overall malignancy fraction was 43%. Notice that there

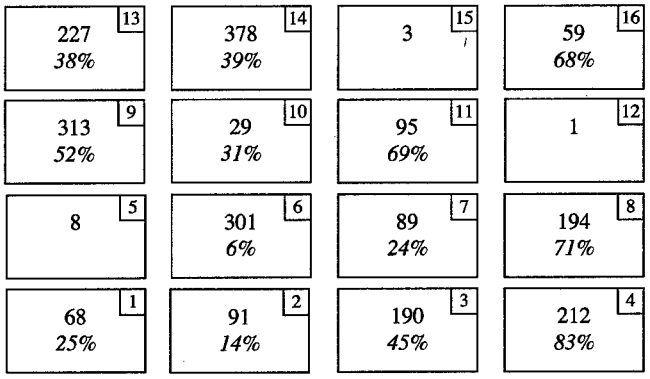

Fig. 2. Index of the neurons in the 4 × 4 map. Each neuron defined a cluster. The number of cases that were mapped to each neuron, i.e. the number of cases in each cluster (normal type), and the fraction of the cases in each cluster that were malignant (italics) is shown. Malignancy fraction data not shown for the clusters with very few cases. Over all, 43% of the cases were malignant. Information regarding the main features of the cases in each cluster is shown in Figs. 4 and 5.

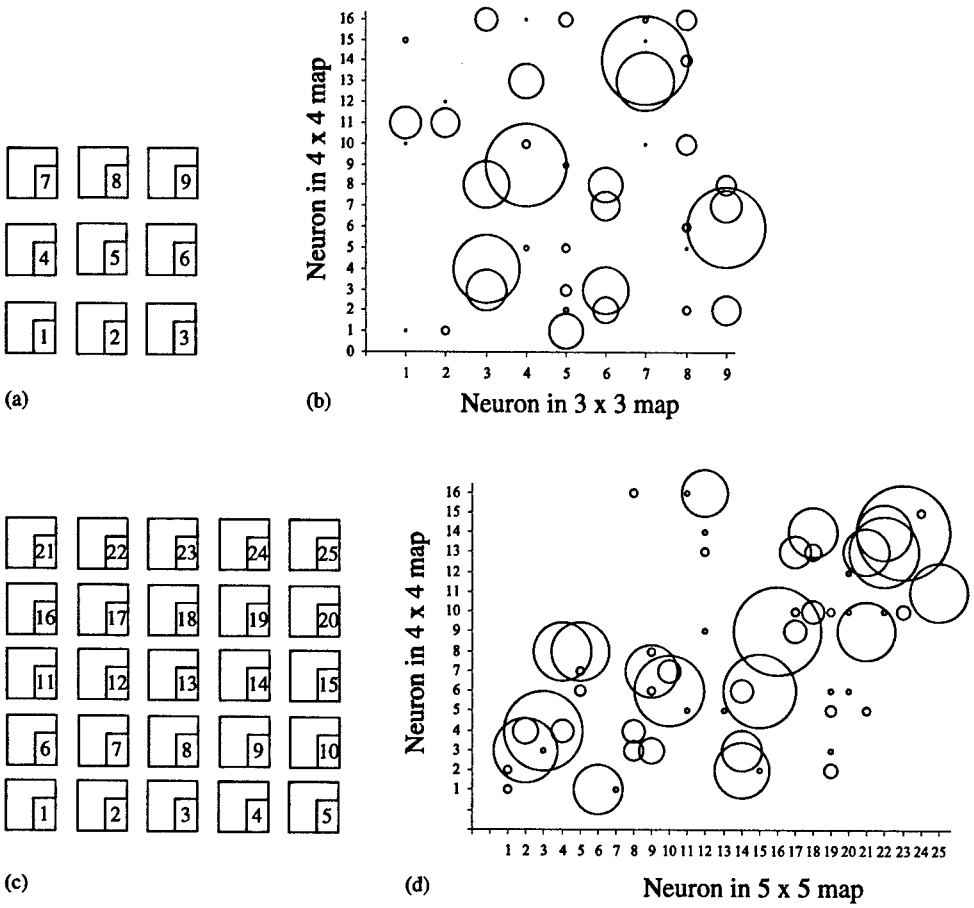

Fig. 3. (a) The index of the neurons in the 3 × 3 map; (b) Comparison of the clusters identified by the 3 × 3 and 4 × 4 SOMs. For each case, the neuron it mapped to was determined for each SOM. The number of cases for each pair of clusters between the two SOMs was plotted, the size of the circle indicates the number of cases. The more large bubbles that are present in such a plot, the more the SOMs agreed on the clustering of the cases. Linear trends (i.e. bubbles lining up along the diagonals) indicate that the same cases are being mapped to the same region in the two SOMs; (c) The index of the neurons in the 5 × 5 map; (d) Comparison of the clusters identified by the 5 × 5 and 4 × 4 SOMs. For each case, the neuron it mapped to was determined for each SOM. The number of cases for each pair of clusters between the two SOMs was plotted; the size of the circle indicates the number of cases. The more large bubbles that are present in such a plot, the more the SOMs agreed on the clustering of the cases. Linear trends (i.e. bubbles lining up along the diagonals) indicate that the same cases are being mapped to the same region in the two SOMs.

is generally a higher incidence of malignant lesions in the clusters on the right-hand side of the map.

Fig. 3 shows the effect that changing the SOM architecture has on the clusters identified. Alternative architectures allow one to vary the number of neurons as well as their topological layout, thus potentially allowing for variations in the complexity of the model. One alternative to a 4 × 4 SOM is a smaller but still square 3 × 3 SOM (Fig. 3a). In Fig. 3b,

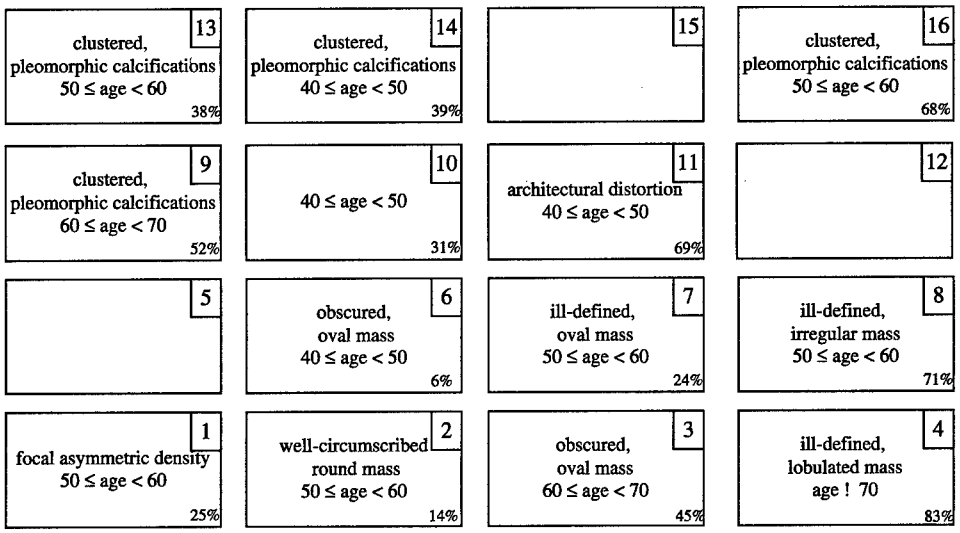

Fig. 4. The cluster profiles generated by the CSNN for the clusters identified by the 4 × 4 SOM (cluster number in upper right corner). A cluster "profile" provides a description of a "typical" case in the cluster. Profiles were not computed for neurons #5, 12, and 15 which had very few cases mapped to them. The percent of the cases that were malignant is shown in the lower right-hand corner; refer to Fig. 2.

the clusters of the 3 × 3 and 4 × 4 SOMs are compared using a bubble plot. For each case, the neuron it mapped to was determined for each SOM. The number of cases for each pair of clusters between the two SOMs was plotted; the size of the circle indicates the number of cases. The more large bubbles that are present in such a plot, the more the SOMs agreed on the clustering of the cases. Similarly, Fig. 3c and 3d show the comparison with a 5 × 5 SOM. Linear trends (i.e. bubbles lining up along the diagonals) indicate that the same cases are being mapped to the same region (e.g. upper right-hand area) in the two SOMs. In addition to square topologies, other layouts were also investigated which utilized approximately the same number of neurons. Comparisons were made to a 2 × 8 SOM and to a three-dimensional SOM of 2 × 3 × 3 neurons, both with approximately the same number of neurons as the 4 × 4 square SOM.

For the 4 × 4 SOM, the cluster profiles generated by the CSNN are shown in Fig. 4. Each cell in the table represents the feature categories that were dominant or most strongly associated with the cases matching that cluster. Profiles were not computed for the clusters with very few cases. The mass cases are distributed over neurons #2, 3, 4, 6, 7, and 8. The profiles of neurons #9, 13, 14, and 16 indicate that those clusters contain microcalcifications. Neuron #1's profile indicates that that cluster is comprised of focal asymmetric densities. Note that the profile for neuron #10 includes only the age variable. The profile for neuron #11 reveals that the lesions in that cluster are architectural distortions.

An alternative approach to generating cluster profiles is to compute summary statistics such as the feature mode (or mean for real-valued features such as age). Fig. 5 shows the mode profiles of the clusters identified by the 4 × 4 SOM. For the most part, there is considerable agreement between the CSNN and mode profiles. Most of the differences

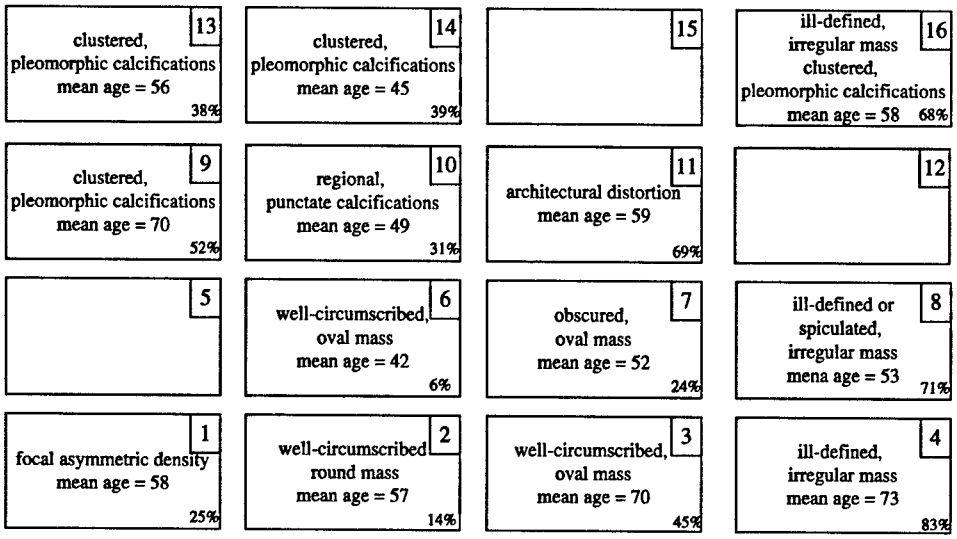

Fig. 5. The cluster profiles generated by the computing the mode the features (mean for age) for the clusters identified by the 4 × 4 SOM (cluster number in upper right corner). A cluster "profile" provides a description of a "typical" case in the cluster. Profiles were not computed for neurons #5, 12, and 15 which had very few cases mapped to them. Features for which the mode value indicated that the feature was absent were omitted (e.g. mass margin = no mass). The percent of the cases that were malignant is shown in the lower right-hand corner; refer to Fig. 2.

correspond to adjacent categories in the features (Table 1) where the CSNN has selected the second most prevalent value for the profile. However, using multiple methods to summarize the clusters may be beneficial. For example, the CSNN profile of neuron #16 (Fig. 4) does not include any mass features yet the feature mode profile (Fig. 5) shows that the mass features are usually non-zero. In fact, inspection of the cases in the cluster defined by neuron #16 reveals that they are calcified masses. Conversely, the CSNN profile for neuron #10 (Fig. 4) includes only the age variable while the mode profile's (Fig. 5) inclusion of values for the calcification variables may be misleading for this small cluster ($N = 29$) where there is little dominance by any single value.

A BP-ANN was trained to predict the biopsy outcome from the BI-RADS^{TM} features and patient age. Fig. 6 shows the ROC curve for the BP-ANN. The SOM can also be used to generate a malignancy prediction [4]. For each case, the prediction was the fraction of the cases that were malignant in the cluster that the case was mapped to by the SOM. For example, if a case belonged to cluster #4 in which 83% of the cases were malignant, then the classifier output for that case would be 0.83. Notice that using this approach limits the number of operating points on the non-parametric ROC curve to the number of clusters with unique malignancy fractions minus one (Fig. 6). The performance at the highest sensitivities was comparable. In particular, at 98% sensitivity the SOM operates with 0.26 ± 0.03 specificity and the BP-ANN operates with 0.25 ± 0.03 specificity ($P = 0.93$).

Fig. 7 lists how the BP-ANN trained on all the cases performs in terms of the BP-ANN's recommendations for follow up instead of biopsy on the subsets identified by the SOM. A

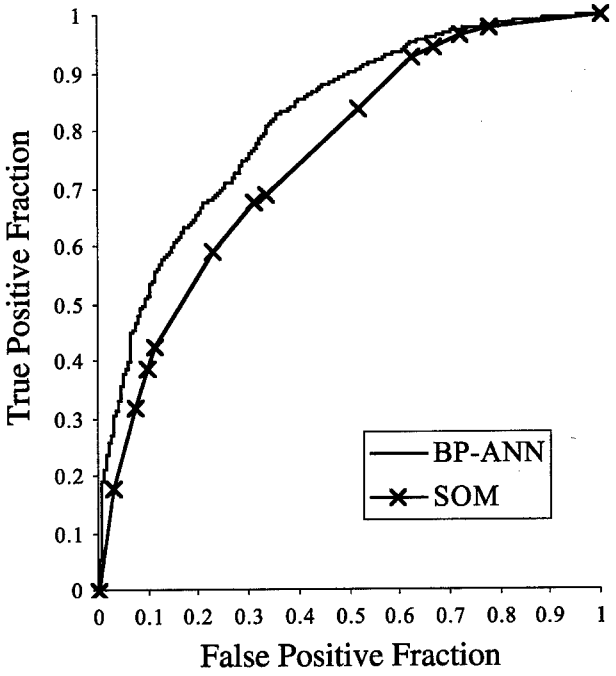

Fig. 6. ROC curves for the BP-ANN and the SOM. For each case, the prediction from the SOM was the fraction of the cases in the cluster it belonged to that were malignant.

threshold was applied to the BP-ANN outputs such that the overall sensitivity was approximately 98% (965/982) with resulting specificity of approximately 24% (303/1276). In other words, 320 cases (303 actual negatives and 17 actual positives) fell below the threshold. These 320 cases that the BP-ANN would have recommended for follow up

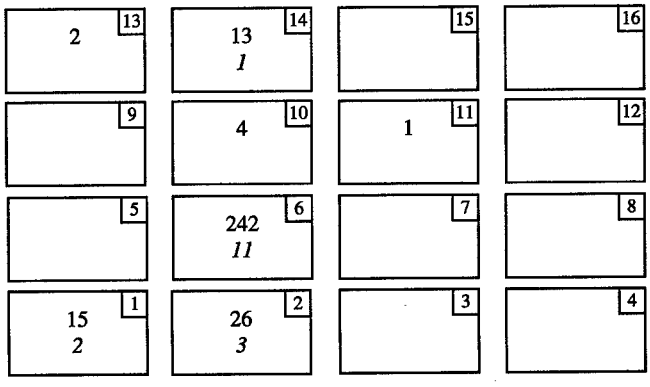

Fig. 7. Comparison of the performance of the BP-ANN trained on all the cases on the clusters identified by the SOM. For each cluster the number of true negatives (normal type) and the number of false negatives (italics) is shown.

are shown in Fig. 7 according to which SOM cluster they belonged. Notice that there is considerable variability in the performance on the clusters. In particular, the majority of the cancers that the BP-ANN would have referred to follow up ($11/17 = 65\%$) and the majority of the benign lesions that the BP-ANN would have spared biopsy ($242/303 = 80\%$) were in the cluster defined by neuron #6.

These interesting results with the cluster defined by neuron #6 suggested that a simple rule-based approach could be valuable. We developed a classification rule based on the cluster profiles (Figs. 4 and 5) of neuron #6 and a classification and regression tree (CART) [3] model for mass cases using the implementation in S-PLUS® (Insightful Corp., Seattle, WA). The classification rule was: if the mass margin was well-circumscribed or obscured and the age was less than 59 years and there were no calcifications, associated findings, or special findings, then do not biopsy, otherwise do biopsy. On the 2258 training cases, this rule gave $961/982 = 98\%$ sensitivity and $336/1276 = 26\%$ specificity. In other words, this rule performed comparably to the BP-ANN with a threshold of 0.1842 ($965/982 = 98\%$ sensitivity, $303/1276 = 24\%$ specificity).

The performance of the BP-ANN and the classification rule developed from data mining were evaluated on the 2177 cases withheld for model validation. On the validation set, the classification rule gave $886/904 = 98\%$ sensitivity and $339/1273 = 27\%$ specificity and the BP-ANN with a threshold of 0.1842 gave $884/904 = 98\%$ sensitivity and $296/1273 = 23\%$ specificity. Thus, both the BP-ANN and the rule-based approach generalized and they performed comparably at this high sensitivity point.

4. Discussion

Considerable variability was seen in the fraction of the cases that were malignant from cluster to cluster. Several clusters had malignancy fractions that were notably different from the fraction of the entire data set (43%). One of the major goals of computer-aided diagnosis of breast cancer is to identify very likely benign cases as candidates for follow up in lieu of biopsy, in order to reduce the number of benign biopsies. Therefore, the clusters with very low malignancy fractions (e.g. neuron #6 with 6% malignant) are dominated by such very likely benign lesions and may be of particular interest for further studies. It is possible to use the clusters and their malignancy fractions directly as a tool for predicting biopsy outcome [4]. For each case, the prediction was the fraction of the cases that were malignant in the cluster that the case was mapped to by the SOM (Fig. 6). For very high sensitivities, this prediction scheme (98% sensitivity, 0.26 ± 0.03 specificity) was competitive with the back-propagation artificial neural network (98% sensitivity, 0.25 ± 0.03 specificity, $P = 0.93$); however, this SOM-based method was not superior to the BP-ANN. The SOM prediction method in conjunction with the CSNN profiling method has the potential advantage that physicians may understand the intuition behind it better than they do the BP-ANN, which is often seen as a "black box". The SOM prediction method, similar to a case-based reasoning system, predicts the probability of malignancy of a new case by reporting the fraction of similar cases that were found to be malignant [7]. The SOM prediction method could also potentially be used in an ensemble of classifiers. If the outputs of two classifiers are not strongly correlated, it is

possible that they could be combined to produce a classifier that is better than either of its component classifiers.

The effects of the changing the SOM architecture were investigated (Fig. 3). As indicated by the presence of large circles in the bubble plots, the SOMs with similar architectures showed substantial agreement in clustering the data. Moreover, the presence of linear trends in the comparisons with the 5×5, 2×8, and $2 \times 3 \times 3$ SOMs suggest that similar SOM architectures result in similar geometric relationships between clusters. These data argue that the clustering is relatively insensitive to the SOM architecture for this problem. It should be noted that this study did not focus on the organization of the clusters into a topological map. Consequently, many of the analyses in this study could have been performed using other clustering algorithms.

Fig. 4 lists the CSNN profiles for the clusters identified with the SOM. The successful separation of a priori known, coarse lesion types (masses, clustered microcalcifications, focal asymmetric densities, and architectural distortions) provided some quality assurance of the clustering. Clusters were further identified within the general group of mass lesions, reflecting different combinations of the mass margin, mass shape, and patient age variables. The cluster profiles that included calcification features showed stratification of the general group of calcification lesions only by patient age and not any of the calcification findings. Notice that while some features may not be considered useful by the CSNN for profiling individual clusters, it is possible that they could be useful to other summarizing techniques or to methods designed to describe the differences between clusters.

An alternative approach to characterizing the clusters is to calculate summary statistics for each of the features. Fig. 5 shows the mode for each of the BI-RADS^{TM} features and the mean of the patient age for each cluster. In general, there is good agreement in the cluster descriptions obtained from these summary plots and the CSNN profiles. However, they are not identical. The most notable differences are for neurons #10 and 16, which show the advantages and disadvantages, respectively, of the fact that the CSNN method inherently includes feature selection.

It may be easier to interpret a CSNN profile, with typically only a few dominant features per cluster, than to interpret as many summary values as there are input findings. Note as well that the CSNN takes into the account interdependencies between the features, while the summary statistics were based on each feature independently. CSNN profiles or summary statistics can be used to quickly sort through the results of a clustering technique, but additional characterization may be appropriate for clusters of particular interest.

Classification based on the SOM was competitive to that achieved by the BP-ANN at high sensitivity levels (Fig. 6). Notable variation in the performance over the clusters identified by the SOM was observed (Fig. 7). This is consistent with our previous work demonstrating performance differences with an a priori partitioning of the data into two broad subgroups of lesions, masses and microcalcifications [17] and suggests that further work should be done to investigate building cluster-specific models. The variation in the BP-ANN performance across the clusters could also influence the ultimate clinical implementation of the decision aid since it may not be useful to apply the BP-ANN to cases similar to those groups of cases for which it always recommended biopsy in the training set. Interestingly, the SOM identified a cluster of mass cases (#6) which accounted

for the majority cases that the BP-ANN would have recommended for follow up rather than biopsy. Recall that the identification of likely benign cases that could be spared biopsy is the goal of such computer-aided diagnosis schemes. This suggests that the SOM clustering and CSNN profiling technique could be used to provide the physician with an alternative description of what the BP-ANN does for certain types of cases. The identification of a single cluster that accounted for the majority of the cases that the BP-ANN would have recommended for follow up also suggests the investigation of rule-based methods to identify relatively simple diagnostic criteria which might be applied to these cases to aid the radiologists in their decision making process. Based on the profiles of the clusters identified by the SOM, we developed a simple classification rule that performed comparably to the BP-ANN (approximately 25% specificity with 98% sensitivity). Moreover, we demonstrated that the classification rule generalized to 2177 cases withheld for model validation.

Acknowledgements

This work was supported in part by Susan G. Komen Breast Cancer Foundation grant DISS0100400, US Army Medical Research and Materiel Command grants DAMD17-02-1-0373 and DAMD17-01-1-0516, and NIH/NCI R29 CA-75547. We would like to thank Brian Harrawood for scientific programming.

References

[1] Baker JA, Kornguth PJ, Lo JY, Williford ME, Floyd Jr. CE. Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon. Radiology 1995;196:817–22.
[2] Bishop CM, Neural networks for pattern recognition. New York: Oxford University Press; 1995.
[3] Breiman L, Friedman J, Olshen R, Stone C, Classification and regression trees. Belmont: Wadsworth International Group; 1984.
[4] Chen D, Chang RF, Huang YL. Breast cancer diagnosis using self-organizing map for sonography. Ultrasound in Med Biol 2000;26:405–11.
[5] Doi K, MacMahon H, Katsuragawa S, Nishikawa RM, Jiang Y. Computer-aided diagnosis in radiology: potential and pitfalls. Eur J Radiol 1999;31:97–109.
[6] Efron B, Tibshirani RJ, An Introduction to the bootstrap. New York, NY: Chapman & Hall; 1993.
[7] Floyd Jr. CE, Lo JY, Tourassi GD. Case-based reasoning computer algorithm that uses mammographic findings for breast biopsy decisions. Am J Roentgenol 2000;175:1347–52.
[8] Giger ML. Computer-aided diagnosis of breast lesions in medical images. Comput Sci Eng 2000;2:39–45.
[9] Heath M, Bowyer KW, Kopans D, Current status of the digital database for screening mammography. In: Karssemeijer N, Thijssen M, Hendriks JH, editors. Digital mammography. Dordrecht: Kluwer Academic Publishers; 1998. p. 457–60.
[10] Hertz J, Anders K, Palmer RG, Introduction to the theory of neural computation. Redwood City, CA: Addison-Wesley; 1991.
[11] Huo Z, Giger ML, Metz CE. Effect of dominant features on neural network performance in the classification of mammographic lesions. Phys Med Biol 1999;44:2579–95.
[12] Huo Z, Giger ML, Vyborny CJ, Wolverton DE, Metz CE. Computerized classification of benign and malignant masses on digitized mammograms: a study of robustness. Acad Radiol 2000;7:1077–84.
[13] Kohonen T, Self-organizing maps. Berlin: Springer; 1995.
[14] Li L, Zheng Y, Zheng L, Clark RA. False-positive reduction in CAD mass detection using a competitive classification strategy. Med Phys 2001;28:250–8.

[15] Lo JY, Baker JA, Kornguth PJ, Floyd Jr. CE. Effect of patient history data on the prediction of breast cancer from mammographic findings with artificial neural networks. Acad Radiol 1999;6:10–5.
[16] Lo JY, Markey MK, Baker JA, Floyd Jr. CE. Cross-institutional evaluation of BI-RADS predictive model for mammographic diagnosis of breast cancer. Am J Roentgenol 2002;178:457–63.
[17] Markey MK, Lo JY, Floyd Jr. CE. Differences between computer-aided diagnosis of breast masses and that of calcifications. Radiology 2002;223:489–93.
[18] Metz CE. Basic principles of ROC analysis. Seminars in Nucl Med 1978;8:283–98.
[19] Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986;21:720–33.
[20] Illustrated breast imaging reporting and data system (BI-RADSTM). Reston, VA: American College of Radiology; 1998.
[21] Rumelhart DE, McClelland JL, Parallel distributed processing: explorations in the microstructure of cognition volume 1: Foundations. Cambridge, MA: MIT Press; 1986.
[22] Rymon R, Zheng B, Chang YH, Gur D. Incorporation of a set enumeration trees-based classifier into a hybrid computer-assisted diagnosis scheme for mass detection. Acad Radiol 1998;5:181–7.
[23] Schaller HN, Constraint satisfaction problems. In: Leondes CT, editor. Optimization techniques. San Diego, CA: Academic Press; 1998. p. 209–48.
[24] Sharkey AJC, Combining artificial neural nets: ensemble and modular multi-net systems. London: Springer; 1999.
[25] Tourassi GD, Markey MK, Lo JY, Floyd Jr. CE. A neural network approach to breast cancer diagnosis as a constraint satisfaction problem. Med Phys 2001;28:804–11.
[26] Vyborny CJ, Giger ML, Nishikawa RM. Computer-aided detection and diagnosis of breast cancer. Radiol Clin North Am 2000;38:725–40.
[27] West D, West V. Model selection for a medical diagnostic decision support system: a breast cancer detection case. Artif Intell Med 2000;20:183–204.
[28] Wu Y, Giger ML, Doi K, Vyborny CJ, Schmidt RA, Metz CE. Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer. Radiology 1993;187:81–7.
[29] Zheng B, Chang Y, Good WF, Gur D. Performance gain computer-assisted detection schemes by averaging scores generated from artificial neural networks with adaptive filtering. Med Phys 2001;28:2302–8.
[30] Zheng B, Chang YH, Gur D. Adaptive computer-aided diagnosis scheme of digitized mammograms. Acad Radiol 1996;3:806–14.
[31] Zheng B, Chang YH, Gur D. Mass detection in digitized mammograms using two independent computer-assisted diagnosis schemes. Am J Roentgenol 1996;167:1421–4.

Computer-aided classification of breast microcalcification clusters: Merging of features from image processing and radiologists

Joseph Y. Lo*[1], Marios Gavrielides[2], Mia K. Markey[3], Jonathan L. Jesneck[1]

[1]Dept. of Radiology, Duke University Medical Center and Dept. of Biomedical Engineering, Duke University, Durham, NC 27710, USA; [2]Artistotle University of Thessaloniki, Thessaloniki Greece; [3]Dept. of Biomedical Engineering, University of Texas at Austin, Austin, TX 78712, USA

ABSTRACT

We developed an ensemble classifier for the task of computer-aided diagnosis of breast microcalcification clusters, which are very challenging to characterize for radiologists and computer models alike. The purpose of this study is to help radiologists identify whether suspicious calcification clusters are benign vs. malignant, such that they may potentially recommend fewer unnecessary biopsies for actually benign lesions. The data consists of mammographic features extracted by automated image processing algorithms as well as manually interpreted by radiologists according to a standardized lexicon. We used 292 cases from a publicly available mammography database. From each cases, we extracted 22 image processing features pertaining to lesion morphology, 5 radiologist features also pertaining to morphology, and the patient age. Linear discriminant analysis (LDA) models were designed using each of the three data types. Each local model performed poorly; the best was one based upon image processing features which yielded ROC area index A_Z of 0.59 ± 0.03 and partial A_Z above 90% sensitivity of 0.08 ± 0.03. We then developed ensemble models using different combinations of those data types, and these models all improved performance compared to the local models. The final ensemble model was based upon 5 features selected by stepwise LDA from all 28 available features. This ensemble performed with A_Z of 0.69 ± 0.03 and partial A_Z of 0.21 ± 0.04, which was statistically significantly better than the model based on the image processing features alone ($p<0.001$ and $p=0.01$ for full and partial A_Z respectively). This demonstrated the value of the radiologist-extracted features as a source of information for this task. It also suggested there is potential for improved performance using this ensemble classifier approach to combine different sources of currently available data.

Keywords: computer-aided diagnosis, breast cancer, BI-RADS, image processing, ensemble classifier

1. INTRODUCTION

1.1 Clinical significance

Mammography is the modality of choice for early detection of breast cancer. Although mammography is very sensitive at detecting breast cancer, its low positive predictive value (PPV) results in biopsy of a large number of benign lesions. Of women with radiographically-suspicious, nonpalpable lesions who are sent to biopsy, only 15 to 34% actually have a malignancy by histologic diagnosis [1,2]. The excessive biopsy of benign lesions raises the cost of mammographic screening [3] and results in emotional and physical burden to the patients, as well as financial burden to society. It is imperative to improve the specificity of breast biopsy by identifying probably benign lesions for short-term follow-up instead of biopsy, while maintaining the very high sensitivity of cancer detection [4,5].

The presence of clustered microcalcifications is one of the most important and sometimes the only sign of cancer on a mammogram [6]. In a recent study from this institution, radiologists demonstrated an interesting dichotomy in performance when asked to assess the likelihood of malignancy among 1468 nearly consecutive mammography cases

* email Joseph.Lo@Duke.edu; home page http://bishop.mc.duke.edu/~jyl

[7]. They performed significantly better as measured by ROC area index (A_Z) for the mass cases (0.94 ± 0.01) compared to the calcifications (0.74 ± 0.02). Similar trends were observed for a variety of statistical and machine learning modeling techniques as well. Another study from a different institution reported similarly low radiologists' performance (A_Z of 0.61) for 104 nearly consecutive microcalcification cases [8].

These studies indicate there is tremendous room for improvement for these calcification cases. If performance can be improved for these challenging cases, overall specificity will in turn be dramatically increased. It should be noted that in clinical practice, the radiologist's task is to recommend whether or not to biopsy, rather than predicting an explicit likelihood of malignancy among lesions already recommended for biopsy. Nevertheless, the fact remains that two-thirds or more of currently biopsied cases are actually benign. In order to improve specificity of breast biopsy, the additional challenge becomes to identify *a priori* more very likely benign cases among those currently referred to biopsy.

1.2 Computer aided diagnosis

It is important to distinguish computer-aided detection vs. computer-aided diagnosis or classification. For computer-aided detection (CAD), a suspicious lesion is detected and localized by some automated computer vision technique such as those in the academic literature [9-11] or one of the currently available commercial systems. The main goal of computer-aided detection is to improve sensitivity by helping radiologists catch disease which might otherwise have been missed. Once the lesion has been detected by radiologists and/or some computer-aided detection device, a computer-aided diagnosis (CADx) system then helps the radiologist to classify that lesion or to make a patient management decision. The main goal of computer-aided diagnosis is typically to help improve specificity, such as by sparing unnecessary benign biopsies.

We propose a classifier to aid in the decision task of whether to biopsy a suspicious lesion or to refer the case to short-term follow-up surveillance. Correct diagnoses have the following implications: very likely benign cases may undergo follow-up instead of biopsy, while the remaining indeterminate cases should undergo biopsy for confirmation by histopathologic diagnosis. Incorrect diagnoses have the following implications: false positive errors (benign lesions misclassified as malignant) may result in an unnecessary biopsy, while false negative errors (malignancies misclassified as benign) may result in delayed diagnosis of an actual cancer. Since the implications for false negative errors far outweigh those of false positives, CADx systems are typically evaluated at operating points corresponding to very high sensitivity.

1.3 Local vs. ensemble models

There have been three major approaches to CADx of the breast biopsy decision task, depending on the source of the input data. The first is to employ image processing techniques which extract features from digitized or digital mammograms [8,12-19]. These fully- or semi-automated CADx systems are not constrained by the limits of human vision, should be more consistent, and have the potential to improve the performance of less experienced radiologists.

The second approach is to rely upon radiologists to interpret the images and manually record findings deemed clinically relevant [20-26]. This approach draws upon the *a priori* knowledge of these radiologists, who can characterize a tremendous amount of image information into a list of succinct, useful findings, such as the standardized lexicon known as the Breast Imaging Reporting and Data System (BI-RADS; American College of Radiology, Reston, VA) [27]. Moreover, such input data is often already available and intuitively meaningful to physicians, which may facilitate eventual clinical acceptance of systems based on such data. The disadvantages are the limits of human vision and knowledge, as well as potential problems arising from intra- and inter-observer variability.

A third approach uses additional information including patient history, clinical, or demographic data. Such data tend to correlate less well with disease, and can often be very subjective in quality and laborious to collect. An exception may be the patient age, which is readily available and was identified as a surprisingly useful adjunct in predicting malignancy in our previous work [28].

We will investigate combining these three sources of data (image processing, BI-RADS, and history) into one ensemble system. An ensemble system uses multiple classifiers to solve a classification problem by training multiple models for the same cases and then combining models' predictions [29]. Simple ensembles of classifiers using voting or averaging to combine their predictions have shown promise in this field [30-33]. The hypothesis here is that an ensemble classifier comprised of information from all three sources of data can significantly outperform models based upon local subsets of data.

2. MATERIALS AND METHODS

2.1 Database

Until recently all major research laboratories reported results based upon private databases. The performance of an algorithm is affected by the characteristics of a database including digitizer choice, pixel size, subtlety of cases, choice of training/testing subsets, and the number of cases in each subset, thus making it almost impossible to compare results reported from different research groups [34]. The establishment of the Digital Database for Screening Mammography (DDSM) [35] allows the possibility of common training and testing data sets for the first time. The DDSM is the largest publicly available database of mammographic data. It contains approximately 2000 screening mammography cases obtained between 1988 and 1999 at several institutions including Massachusetts General Hospital, Wake Forest University School of Medicine, Sacred Heart Hospital, and Washington University in St. Louis School of Medicine.

For this pilot study, we specified a patient selection criteria to provide a reasonable number of cases while keeping methodological and statistical issues as simple as possible. From cases with definitive pathology outcome (i.e. not "unproven" or "benign, no call back"), we randomly selected 292 cases which were digitized by the Howtek digitizer (which had the most cases compared to the other 2 types of digitizers). Each case had only one cluster recorded in the truth file, and that cluster was successfully segmented by our existing automated detection algorithm which has been described in detail previously [36,37]. All image processing was performed only on the medio-lateral oblique (MLO) view of the breast containing the lesion, thus obviating any problems due to per-case vs. per-patient sampling and performance analysis.

2.2 Computer-extracted features

For each case, we used the aforementioned detection technique [36,37] as the front end to localize clusters and segment individual calcifications within those clusters. This fully automated detection scheme consisted of three main processing steps:

(1) Pre-processing. The breast region was segmented and its high frequency content was enhanced by unsharp masking.
(2) Segmentation of individual calcifications. Individual microcalcifications were segmented using local histogram analysis on small, overlapping regions of interest (ROIs). Each histogram was modeled as a possible bimodal distribution of bright calcifications on a darker background. Histogram features were extracted and then merged using a back-propagation artificial neural network (BP-ANN) classifier [38-40] to determine whether each ROI contained a calcification.
(3) Cluster classification. The calcifications were clustered using a nearest neighbor algorithm. Features were extracted describing each cluster and then merged using another BP-ANN classifier to reduce the number of false positive clusters.

For each cluster, 22 image processing features were calculated. These consisted of the number of calcifications, logarithm of that number, total area of all calcifications, logarithm of that area, and the mean and standard deviation of each of the following nine morphological features: calcification distance, number of overlaps (resulting from the overlapping ROIs in histogram analysis), calcification area, compactness, central moment, Fourier descriptor, eccentricity, spread, and orientation.

A region from a sample case containing a malignant calcification cluster and the corresponding detection output are shown in Figure 1.

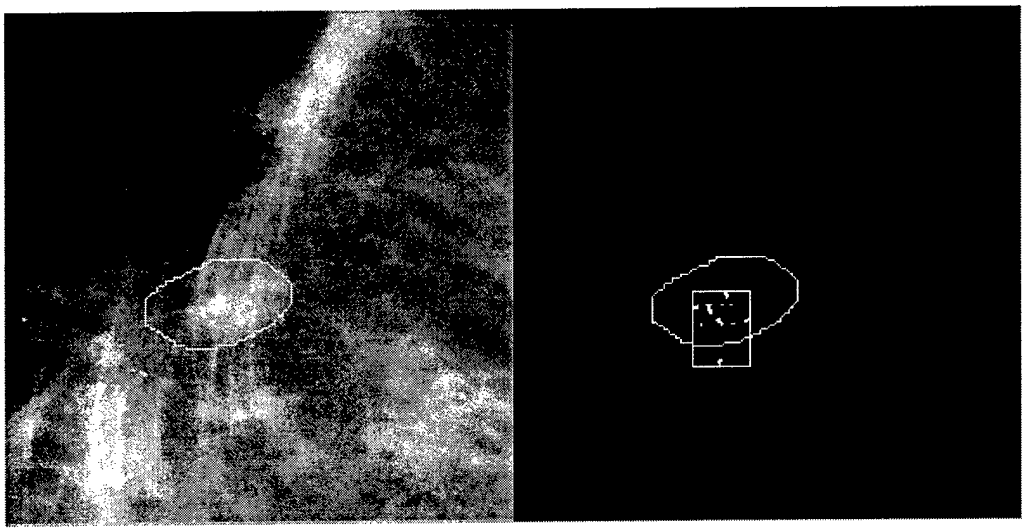

Figure 1. Sample detection output for malignant cluster (left: case1108, left CC, cluster outlined by experienced radiologist, right: true positive detected cluster bounded by rectangle).

Table 1. BI-RADS mammographic features and numeric encoding

Calc. Distribution		**Mass Margin**	
no calcifications	0	no mass	0
diffuse	1	well circumscribed	1
regional	2	microlobulated	2
segmental	3	obscured	3
linear	4	ill-defined	4
clustered	5	spiculated	5
Calc. Morphology		**Mass Shape**	
no calcifications	0	no mass	0
milk of calcium-like	1	round	1
eggshell or rim	2	oval	2
skin	3	lobulated	3
vascular	4	irregular	4
spherical or lucent centered	5		
suture	6	**Associated Findings**	
coarse	7	none	0
large rod-like	8	skin lesion	1
round	9	hematoma	2
dystrophic	10	post surgical scar	3
punctate	11	trabecular thickening	4
indistinct	12	skin thickening	5
pleomorphic	13	skin retraction	6
fine branching	14	nipple retraction	7
		axillary adenopathy	8
		architectural distortion	9

2.3 Human-extracted features

For each case, we also extracted 5 BI-RADS features and the patient age from the database. The BI-RADS features were: calcification morphology, calcification distribution, mass shape, mass margin, and associated findings. Note that the two mass findings occur because these calcification cases were defined as those with the presence of calcification findings. In some cases, an associated mass was also present. The text labels for each BI-RADS feature were translated into numeric values using a rank ordering system shown below in Table 1 which we have used previously in developing models with this type of data [28,41]. In cases where a feature was described by multiple values, such as if there were two values for the calcification distribution, the greatest value corresponding to the highest likelihood of malignancy was used.

2.4 Statistical Sampling and Measurements

Due to the relatively low number of cases available, all modeling was performed using linear discriminant analysis (LDA) using SAS software (SAS Inc., Cary NC) with round robin sampling. A_Z and partial A_Z above sensitivity of 90% were calculated using LABROC4 and compared using CLABROC (both modified by Charles Metz, University of Chicago, to provide partial A_Z calculations). The partial A_Z was used to characterize the more clinically relevant high sensitivity sub-region of the ROC curve, which emphasizes the far greater cost of a missed cancer compared to an unnecessary benign biopsy [42].

3. RESULTS

The results are summarized in Table 2 and Figure 2 below. Each row represents the performance of a model based upon a combination of one or more of the three sources of data, which have been color coded: A) blue for image processing features, B) pink for BI-RADS, and C) green for the sole history feature of age. The columns labeled as A, B, and C indicate that the model on that row used some or all of the features from that source of data.

None of the 3 sources of patient data by itself provided much useful information, as shown in rows *A, B, and C* all with $A_Z < 0.6$. There were however interesting improvements when these data were combined together. For example, on row *D*, the addition of just age (which by itself performed close to chance) significantly improved performance over the 5 BI-RADS features alone in row *B* ($p<0.001$ for both full and partial A_Z). On row *E*, the further addition of the 22 image processing features, i.e., using all 28 available features, did not improve performance compared to row *D*. On row *F*, when stepwise LDA was used to reduce those 28 total features to just 5, however, that yielded the best performance of all at A_Z of 0.69 and partial A_Z of 0.21. This final 5-feature model is significantly better than using only the 22 image processing features (row *F* vs. row *A*) for both full and partial A_Z ($p<0.001$ and $p=0.01$ respectively). The final 5-feature model was not however significantly better the 6 human-extracted features (row *F* vs. row *D*) for either full or partial A_Z ($p=0.07$ and $p=0.15$ respectively). Those final 5 features were (in order of descending significance): BI-RADS calcification distribution, mean central moment, mean eccentricity, BI-RADS mass margin, and BI-RADS calcification morphology.

Table 2. Performance of LDA models with different feature combinations from 292 DDSM cases.

Feature combination	A	B	C	A_Z	Partial A_Z
A) 22 image processing only	■			0.59 ± 0.03	0.08 ± 0.03
B) 5 BI-RADS by radiologist only		■		0.58 ± 0.03	0.07 ± 0.02
C) Age alone			■	0.52 ± 0.03	0.05 ± 0.02
D) All 6 human extracted (5 BI-RADS + age)		■	■	0.66 ± 0.03	0.13 ± 0.03
E) All 28 features (22 image + 5 BI-RADS + age)	■	■	■	0.65 ± 0.03	0.11 ± 0.03
F) Stepwise selection of top 5 features from all 28	■	■		0.69 ± 0.03	0.21 ± 0.04

The ROC curves for rows *A*, *D*, and *F* are plotted below in Figure 2. As described above, row *A* with the image processing features only performed poorly (shown by red line), and the final ensemble including contributions from BI-

RADS features improved that performance significantly (row *F* shown by gold line). The human extracted features only from row *D* were intermediate between those two curves.

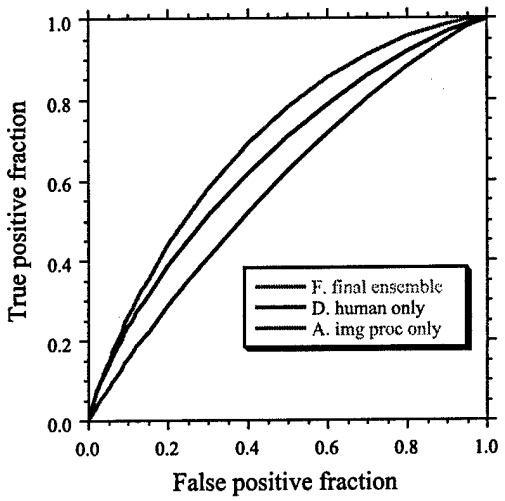

Figure 2. ROC Curves for Ensemble vs. Local Models

4. DISCUSSION

These results suggest several interesting trends. Local models based upon these image processing features or the BI-RADS features each performed comparably (in fact comparably poorly). The addition of age to the BI-RADS features significantly improved performance, which is consistent with our previous experience with these human-only models. The further addition of image processing features improved performance even further, albeit not significantly. That may change with either more cases or better image processing features. For example, there are many other morphological features not used here, as well as several different categories of texture features which have been shown to be very useful for this particular task [14].

Intriguingly, the feature-reduced model did not include age as one of its remaining features. Apparently the significance of age was much decreased in the presence of these image processing and BI-RADS features, a fact that warrants further investigation. The final ensemble model based upon 2 image processing and 3 BI-RADS features did significantly outperform one based upon just the 22 available image processing features, supporting once again the value of these BI-RADS findings in this decision task.

It should be noted that although the trends support the value of building ensemble models for this data, these ROC performance values were still quite poor. The best A_Z was only 0.69 and the best partial A_Z 0.21, corresponding to the average specificity over the range of sensitivities from 0.90 to 1.00. Given the equally poor performance of radiologists for calcification cases, however, there is great potential for improvement. In the end, the most important test in the future will be to assess whether radiologists can use such models to improve their clinical performance.

ACKNOWLEDGMENTS

This work was supported in part by the following grants: NIH / NCI CA92573, DOD Breast Cancer Research Program DAMD17-02-1-0373, and Susan G. Komen Breast Cancer Foundation DISS 2000 729 and DISS 01 00400.

REFERENCES

1 Knutzen AM, and Gisvold JJ, "Likelihood of malignant disease for various categories of mammographically detected, nonpalpable breast lesions," *Mayo Clin Proc* 68, 454-460 (1993).

2 Kopans DB, "The positive predictive value of mammography," *AJR* 158, 521-526 (1992).

3 Cyrlak D, "Induced costs of low-cost screening mammography," *Radiology* 168, 661-3 (1988).

4 Varas X, Leborgne F, and Leborgne JH, "Nonpalpable, probably benign lesions: role of follow-up mammography," *Radiology* 184, 409-14 (1992).

5 Sickles EA, "Management of probably benign breast lesions," *Radiol Clin North Am* 33, 1123-30 (1995).

6 Bassett LW, and Gambhir S, "Breast imaging for the 1990s," *Sem Onc* 18, 80-86 (1991).

7 Markey MK, Lo JY, and Floyd CE, Jr, "Differences in computer aided diagnosis of breast cancer: masses vs. calcifications," in Chicago 2000: World Congress on Medical Physics and Biomedical Engineering, Chicago, IL (Chicago, IL, 2000).

8 Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Giger ML, and Doi K, "Improving breast cancer diagnosis with computer-aided diagnosis," *Acad Radiol* 6, 22-33 (1999).

9 Zheng B, Chang YH, Wang XH, Good WF, and Gur D, "Feature selection for computerized mass detection in digitized mammograms by using a genetic algorithm," *Acad Radiol* 6, 327-32 (1999).

10 Qian W, Clarke LP, Song D, and Clark RA, "Digital mammography: hybrid four-channel wavelet transform for microcalcification segmentation," *Acad Radiol* 5, 354-64 (1998).

11 Qian W, Li L, Clarke L, Clark RA, and Thomas J, "Digital mammography: comparison of adaptive and nonadaptive CAD methods for mass detection," *Acad Radiol* 6, 471-80 (1999).

12 Huo A, Giger ML, and Vyborny CJ, "Analysis of Computer-Aided Diagnosis on Radiologists' Performance Using an Independent Database," *Proc SPIE* 4324, 45-51 (2001).

13 Chan HP, Sahiner B, Helvie MA, Petrick N, Roubidoux MA, Wilson TE, Adler DD, Paramagul C, Newman JS, and Sanjay-Gopal S, "Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study," *Radiology* 212, 817-27 (1999).

14 Chan HP, Sahiner B, Lam KL, Petrick N, Helvie MA, Goodsitt MM, and Adler DD, "Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces," *Med Phys* 25, 2007-19 (1998).

15 Thiele DL, Kimme-Smith C, Johnson TD, McCombs M, and Bassett LW, "Using tissue texture surrounding calcification clusters to predict benign vs malignant outcomes," *Med Phys* 23, 549-55 (1996).

16 Schmidt F, Sorantin E, Szepesvari C, Graif E, Becker M, Mayer H, and Hartwagner K, "An automatic method for the identification and interpretation of clustered microcalcifications in mammograms," *Phys Med Biol* 44, 1231-43 (1999).

17 Giger ML, Al-Hallaq H, Huo Z, Moran C, Wolverton DE, Chan CW, and Zhong W, "Computerized analysis of lesions in US images of the breast," 6, 665-74 (1999).

18 Chang RF, Kuo WJ, Chen DR, Huang YL, Lee JH, and Chou YH, "Computer-aided diagnosis for surgical office-based breast ultrasound," *Archives of Surgery* 135, 696-9 (2000).

19 Chen D, Chang RF, and Huang YL, "Breast cancer diagnosis using self-organizing map for sonography," *US in Med and Biol* 26, 405-11 (2000).

20 Lo JY, Baker JA, Kornguth PJ, Iglehart JD, and Floyd CE, Jr, "Predicting breast cancer invasion with artificial neural networks on the basis of mammographic features," *Radiology* 203, 159-163 (1997).

21 Kahn CE, Jr., Roberts LM, Shaffer KA, and Haddawy P, "Construction of a Bayesian network for mammographic diagnosis of breast cancer," *Computers in Biology & Medicine* 27, 19-29 (1997).

22 Baker JA, Kornguth PJ, Lo JY, Williford ME, and Floyd CE, Jr, "Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon," *Radiology* 196, 817-822 (1995).

23 Floyd CE, Jr, Lo JY, Yun AJ, Sullivan DC, and Kornguth PJ, "Prediction of breast cancer malignancy using an artificial neural network," *Cancer* 74, 2944-2948 (1994).

24 Wu Y, Giger ML, Doi K, Vyborny CJ, Schmidt RA, and Metz CE, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," *Radiology* 187, 81-87 (1993).

25 D'Orsi CJ, Getty DJ, Swets JA, Pickett RM, Seltzer SE, and McNeil BJ, "Reading and decision aids for improved accuracy and standardization of mammographic diagnosis," *Radiology* 184, 619-622 (1992).

26 Getty DJ, Pickett RM, D'Orsi CJ, and Swets JA, "Enhanced interpretation of diagnostic images," *Invest Radiol* 23, 240-252 (1988).

27 BI-RADS. American College of Radiology. *American College of Radiology Breast Imaging - Reporting and Data System (BI-RADS) 3rd ed.* 1998.

28 Lo JY, Baker JA, Kornguth PJ, and Floyd CE, Jr, "Effect of patient history data on the prediction of breast cancer from mammographic findings with artificial neural networks," *Acad Radiol* 6, 10-15 (1999).

29 Sharkey AJC, Ed., *Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems*, (Springer-Verlag, 1999).

30 Li L, Zheng Y, Zheng L, and Clark RA, "False-positive reduction in CAD mass detection using a comptetive classification strategy," *Med Phys* 28, 250-258 (2001).

31 Rymon R, Zheng B, Chang YH, and Gur D, "Incorporation of a set enumeration trees-based classifier into a hybrid computer-assisted diagnosis scheme for mass detection," *Acad Radiol* 5, 181-7 (1998)(98183485).

32 Zheng B, Chang YH, and Gur D, "Mass detection in digitized mammograms using two independent computer-assisted diagnosis schemes," *AJR* 167, 1421-4 (1996).

33 Zheng B, Chang Y, Good WF, and Gur D, "Performance gain computer-assisted detection schemes by averaging scores generated from artificial neural networks with adaptive filtering," *Med Phys* 28, 2302-2308 (2001).

34 Giger ML, Karssemeijer N, and Aramato SG, III, "Guest editorial computer-aided diagnosis in medical imaging," 20, 1205-1208 (2001).

35 Heath M, Bowyer KW, and Kopans D, "Current status of the Digital Database for Screening Mammography," *Digital Mammography*, edited by N. Karssemeijer, M. Thijssen, and J. Hendriks. (Kluwer Academic Publishers, 1998) 457-460.

36 Gavrielides MA, Lo JY, Vargas-Voracek R, and Floyd CE, Jr, "Segmentation of suspicious clustered microcalcifications in mammograms," *Med Phys* 27, 13-22 (2000).

37 Gavrielides MA, Lo JY, and Floyd CE, Jr, "Parameter optimization of a computer-aided diagnosis scheme for the segmentation of microcalcification clusters in mammograms," *Med Phys* 29, 475-483 (2002).

38 Rumelhart DE, and McClelland JL, Ed., *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, (The MIT Press, Cambridge, Massachusetts, 1986).

39 Bishop CM, *Neural Networks for Pattern Recognition*, (Oxford University Press, 1995).

40 Hertz J, Anders K, and Palmer RG, *Introduction to the Theory of Computation, Santa Fe Institute Studies in the Science of Complexity* (Addison-Wesley, 1991).

41 Lo JY, Markey MK, Baker JA, and Floyd CE, Jr, "Cross-institutional evaluation of BI-RADS predictive model for mammographic diagnosis of breast cancer," *AJR* 178, 457-463 (2002).

42 Jiang Y, Metz CE, and Nishikawa RM, "A receiver operating characteristic partial area index for highly sensitive diagnostic tests," *Radiology* 201, 745-750 (1996).